Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

FOUNDATIONS FOR THE REPRESENTATION OF DIVERSITY, EVOLUTION, OPINION AND BIAS

Fausto Giunchiglia, Vincenzo Maltese, Devika Madalli, Anthony Baldry, Cornelia Wallner, Paul Lewis, Kerstin Denecke, Dimitris Skoutas, Ivana Marenzi

November 2009

Technical Report # DISI-09-063

Also: delivery D1.1 of the Living Knowledge EU FET project

LivingKnowledge

LivingKnowledge European Project

- Preliminary report -

Foundations for the representation of diversity, evolution, opinion and bias

---

Document data:	Reference persons:
09.09.30.v01.r02	Fausto Giunchiglia (UniTn), Vincenzo Maltese (UniTn), Devika Madalli (ISI), Anthony Baldry (UniPv), Cornelia Wallner (SORA), Paul Lewis (SOTON), Kerstin Denecke (LUH), Dimitris Skoutas (LUH), Ivana Marenzi (LUH),
	Reviewers: Gerhard Weikum (MPG), Marc Spaniol (MPG).

Index:

Revision History: 3
1. Introduction 4
2. The use case, topic selection and relevant queries 6
3. Analysis of opinion, bias, diversity and evolution 7
3.1. Opinion and bias 7
3.2. Diversity 7
3.3. Knowledge Evolution 8
4. Technological approach to the solution 9
5. Interdisciplinary contributions to the solution 11
6. State of the art 13
6.1. Generic feature extraction and clustering tools: an introduction 13
6.2. Available tools for Text Analysis 16
6.3. Available tools for Image Analysis 20
6.4. Temporal knowledge 26
7. Contributions to the solution from the social and political sciences 29
7.1. Methodological background 29
7.2. Relevant research questions 29
7.3. Levels of analysis 30
7.4. Issues relevant for the chosen topic 31
8. Contributions to the solution from semiotics 35
8.1. A multimodal semiotics standpoint for migration in Websites 35
8.2. Macro-strategies to detect opinions and bias in Websites 43
8.3. Conclusion 46
9. Contributions to the solution from Library and Information Science 47
9.1. Introduction to Faceted Classification 47
9.2. Facet Analysis and Faceted Classification 47
9.3. Basic steps in the construction of a Faceted Classification 48
9.4. Classification of documents according to Faceted Classification 49
9.5 Conclusion 51
10. Interplay between technologies and methodologies 53
References 58
Appendix A: An example of Media Content Analysis 70
Appendix B: Questions which can be answered by the MCA Web Browser 77
Appendix C: Example of the application of MCA Web Browser 80
Appendix D: Glossary of the terms used in Semiotics 83
Appendix E: Glossary of the terms used in faceted approaches 87
Appendix F: Glossary of the terms used in Media Content Analysis 90

Revision History:

Revision	Date (YY/MM/DD)	Author	Description of Changes
v00.r01	09.02.19	Vincenzo Maltese, Fausto Giunchiglia	Document created.
v00.r02	09.03.06	Vincenzo Maltese, Fausto Giunchiglia	Document updated after WP1 meeting on 09.03.06
V00.r03	09.03.19	Cornelia Wallner, Christoph Hofinger	Adding draft version of motivating example from SORA and dimensions of diversity from SORA
V00.r04	09.05.26	Vincenzo Maltese, Fausto Giunchiglia	Integrating all the methodological contributions (SORA, ISI, PAVIA) and adding the technological implications.
V00.r05	09.06.29	Anthony Baldry, Vincenzo Maltese	Semiotic part revisited.
V00.r06	09.07.29	Vincenzo Maltese	The following parts have been added: (1) the role of images; (2) Analysis of opinion, bias and diversity.
V00.r07	09.08.25	Vincenzo Maltese, Gerhard Weikum, Paul Lewis	Knowledge evolution section added. Technological sections added.
V00.r075	09.09.01	Vincenzo Maltese, Richard Johansson	The interplay section has been extended.
V00.r08	09.09.07	Vincenzo Maltese, Fausto Giunchiglia, Anthony Baldry, Cornelia Wallner, Paul Lewis, Dimitris Skoutas, Devika Madalli.	The introduction and the structure of the document have been changed. The whole text has been reviewed.
V01.r00	09.09.17	Vincenzo Maltese, Fausto Giunchiglia, Anthony Baldry, Cornelia Wallner, Paul Lewis, Devika Madalli.	Document ready for the review.
V01.r01	09.09.28	Gerhard Weikum, Marc Spaniol,	Document reviewed.
V01.r02	09.09.30	Vincenzo Maltese, Fausto Giunchiglia, Anthony Baldry, Cornelia Wallner, Devika Madalli, Anand Pandey, Paul Lewis.	Glossaries and document further revised in particular taking into account feedback provided by the reviewers.

1. Introduction

One of the aims of the LivingKnowledge project is to bring a new quality into search and knowledge management technology, by making opinions, bias, diversity and evolution more tangible and more digestible. In order to capture diversity in knowledge, we believe that developing more powerful tools for extracting information from both text and non-text media are keys to the success of the project.

The Web with its abundant amount of contents that continuously cumulate is an excellent example of diversity. It is widely agreed that knowledge is strongly influenced by the diversity of context, mainly cultural, in which it is generated. Thus, while it may be appropriate to say that (some kinds of) cats and dogs are food in some parts of China, Japan, Korea, Laos and the Philippines, this is unlikely to be the case in the rest of the world [223]. A similar example is provided in Fig. 1. Sometimes, it is not just a matter of diversity in culture, viewpoints or opinion, but rather a combination of different perspectives and goals. In fact, knowledge useful for a certain task, and in a certain environment, will often not be directly applicable given other circumstances, and will thus require adaptation. Hence, there is a pressing need to find effective ways of dealing with such complexity, especially in terms of scalability and adaptability in data and knowledge representation.

As first advocated in [117], we are firmly convinced that diversity in knowledge should not be avoided, as often happens in approaches where, at design time, a global representation schema is proposed. We rather see diversity in knowledge as a key feature, our goal being to develop methods and tools leading to effective design by harnessing, controlling and using the effects of emergent knowledge properties. Using these tools, new knowledge can be obtained by adapting existing knowledge but respecting the not entirely predictable process of knowledge evolution and/or aggregation. We envisage a future where developing diversity-aware navigation and search applications will become increasingly important. These applications will automatically classify and organise opinions and bias in order to produce more insightful, better organised, aggregated and easier-to-understand output. This can be accomplished by detecting and differentiating between, what we call, diversity dimensions. This explains our adoption of a highly interdisciplinary approach that brings together expertise from a wide range of disciplines: sociology, philosophy of science, cognitive science, library and information science, semiotics and multimodal information theory, mass media research, communication, natural language processing and multimedia data analysis (among others). The proposed solution is based on the foundational notion of context and its ability to localise meaning, and the notion of facet, as from library science, and its ability to organise knowledge as a set of interoperable components (i.e. the facets).

The purpose of this report is to provide the foundations for the representation and management of diversity, bias and evolution in knowledge. The rest of the report is organised as follows. Section 2 presents the topic selected as our use case and some relevant queries that we want the future system to answer. Section 3 introduces the key notions of opinion, bias, diversity and evolution. Section 4 and Section 5 describe the proposed technological solution and briefly introduce the interdisciplinary contributions to it, respectively. Section 6 mainly summarizes the state of the art as regards available technologies. From Section 7 to 9 the methodologies which actually contribute to the solution are extensively described, namely Media Content Analysis (from a socio-political perspective), Multimodal Analysis (from Semiotics) and Facet Analysis (as used in Library and Information Science for Knowledge Organization purposes), respectively. Section 10 mainly describes current cross research activities between technical and methodology partners and outlines future work. A set of final appendixes are provided. They contain examples of the application of the methodologies presented (especially in terms of document annotation) and glossaries of the relevant technical terms used.

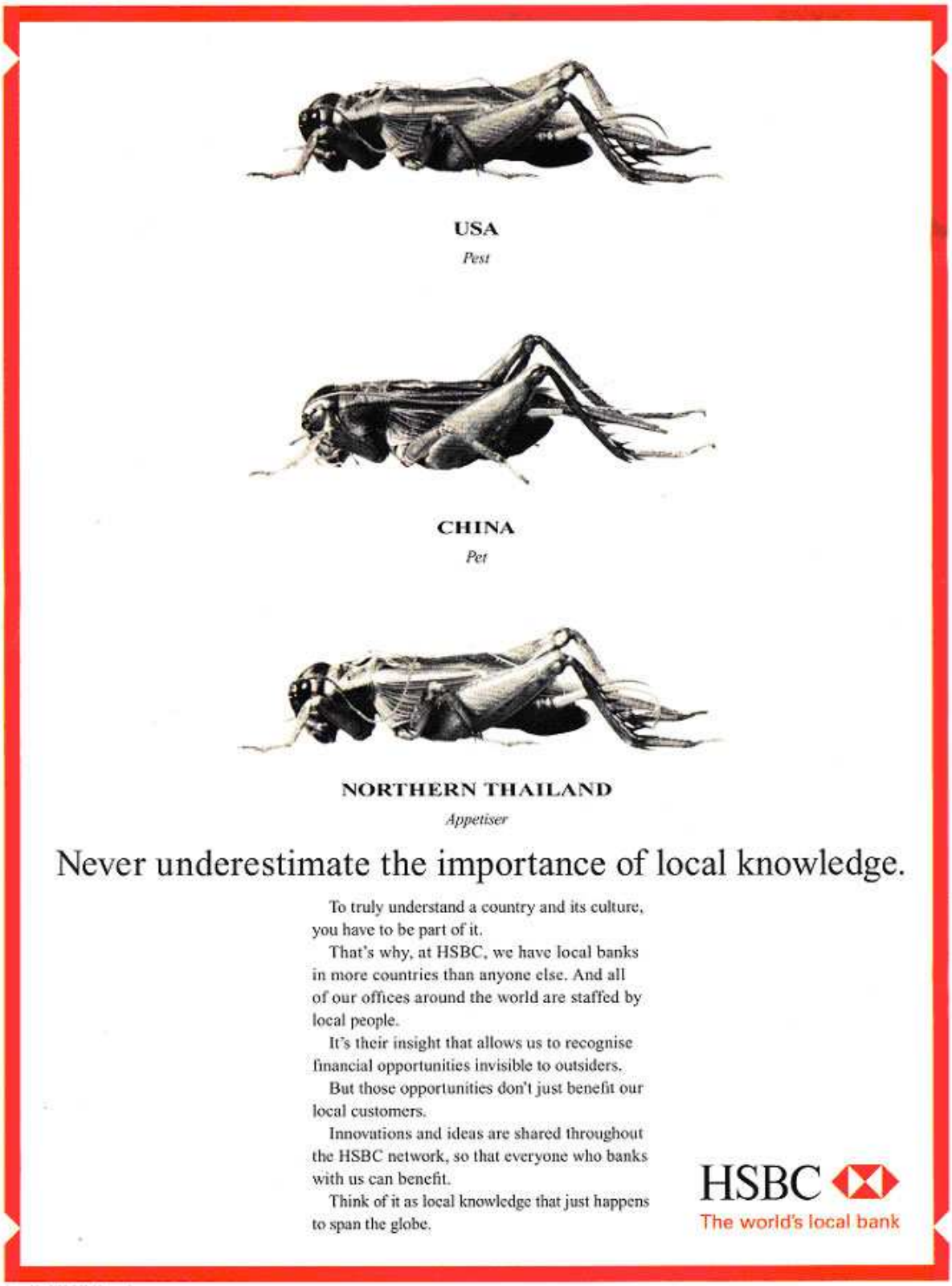

Fig. 1 – A clear example of cultural diversity

2. The use case, topic selection and relevant queries

For the purposes of the project, we decided to select the topic "European elections: migration, xenophobia, integration" as our use case. This topic (and others[1]) will be used in this report to illustrate the project requirements and the individual methodologies and technologies contributing to the solution. In particular, our focus is on the identification and management of diversity in knowledge and its evolution in time, with particular emphasis given to opinion and bias detection. We provide here some examples of relevant queries that we want the system to answer:

- What are the main [*concepts, people, parties, countries, dates, resolutions, etc.*] related to integration?
- Which of these [*concepts, people, parties, countries, dates, resolutions, etc.*] are most [*controversial, accepted, subjective, biased, etc.*]?
- What are the main parties discussing integration in a *negative, positive, controversial, etc.*] context?
- Which parties have changed their discourse on integration (i.e. from positive to negative)?
- Which politicians have changed their discourse on integration (i.e. from positive to negative)?
- Which periods of time are most important vis-à-vis integration, and how are other events correlated to these periods?

Some other typical descriptive research questions (from the socio-political perspective) include:

- What topics occur to what extent in the mediated discourse [*on integration, on the EU-Greens, etc.*]?
- What actors in what roles are present in the mediated discourse on integration?
- What patterns of interpretation (frames) occur in the mediated discourse [*on integration, on the EU-Greens, etc.*]?
- What are the main topics, [*actors and countries, arguments, frames* related *to integration, to the EU-Greens etc.*]?
- What are the main [*politicians, parties, organisations etc.*] discussing [*integration, the EU-Greens, etc.*] in a [*negative, positive, neutral*] context?
- What [*politicians, parties, organisations etc.*] have changed their discourse on integration (i.e. from positive to negative)?
- Which periods of time are most important vis-à-vis integration, and what other events are correlated to these periods?
- How developed is the discursive character of statements from different [*politicians, parties, organisations etc.*]? In terms of an input-output analysis (official material and press releases in comparison to mass media content, forums, blogs):
- How are Green party ideas (as a case) reported in official sites penetrating media, blogs, and forums?
- What differences exist between input (communication on Green Websites, press releases) in as compared with output (discussion in media, blogs, forums)?

In order to conduct our tests, certain documents about the specific topic will be identified. They will constitute the document corpus which is an integral part of the LivingKnowledge testbed. It will contain text and images. We have decided to cover a spectrum of "authority/bias" sources such as Government Websites, official Websites (mainly parties and companies), news, blogs and forums[2].

[1] To illustrate the notions of facet and facet analysis we frequently prefer to use examples from the Medicine domain. This is mainly because it directly maps to the literature we refer to.

[2] Refer to WP6 activities and reports for details.

3. Analysis of opinion, bias, diversity and evolution

The purpose of this section is to introduce the basic notions of opinion, bias, diversity and evolution. In particular, we illustrate how we see them as being closely intertwined.

3.1. Opinion and bias

We define an opinion as follows:

> **Opinion.** *An opinion is a statement, i.e. a minimum semantically self-contained linguistic unit, asserted by at least one actor, called the opinion holder, at some point in time, but which cannot be verified according to an established standard of evaluation. It may express a view, attitude, or appraisal on an entity. This view is subjective, with positive/neutral/negative polarity (i.e. support for, or opposition to, the statement).*

By 'entity' we mean something that has a distinct, separate existence, not necessarily a material existence; it may be a concrete object or an abstract concept. In the sentence "President Obama said that police in Cambridge, Massachusetts, 'acted stupidly' in arresting a prominent black Harvard professor", the opinion holder is *President Obama*, the statement is *police acted stupidly* which expresses an opinion with *negative* polarity. We then define bias as follows:

> **Bias.** *Bias is the degree of correlation between (a) the polarity of an opinion and (b) the context of the opinion holder.*

We thus see bias as a linking device. The polarity of an opinion is the degree to which a statement is positive, negative or neutral. The context may refer to a variety of factors, such as ideological, political, or educational background, ethnicity, race, profession, age, location, or time. Bias is potentially measurable directly in terms of a scale for this correlation e.g. measuring the minority/majority status of opinions in different contexts, particularly in relation to cultural diversity [202][203][204]. For example, by asking the question *What proportions of conservatives, liberals and socialists favoured integration of Turkey into the EU in 1999, 2004 and 2009?* we begin to realise, in a scalable way, whether the polarity of an opinion is correlated with the particular context of the opinion holders and, indeed, whether changes in bias occur over time.

3.2. Diversity

We define diversity in relation to definitions that have emerged from Media Content Analysis [203], [205], [206] as follows:

> **Diversity.** *Diversity is the co-existence of contradictory opinions and/or statements (some typically non-factual or referring to opposing beliefs/opinions).*

There are various forms and aspects of diversity:

- The existence of opinions with different polarity about the same entity, e.g., at different times
- Diversity of themes, speakers, arguments, opinions, claims and ideas
- Diversity of norms, values, behaviour patterns, and mentalities
- Diversity in terms of geographical (local, regional, national, international, global focus of information), social (between individuals, between and within groups), and systemic (organizational and societal) aspects in media content
- Static (at one point in time) and dynamic (long-term) diversity
- Internal diversity (within one source) and external diversity (between sources)

Generally speaking, the following **dimensions of diversity** can be distinguished, both in texts and images:

- Diversity of sources (multiplicity of sources of texts and images)

- Diversity of resources (e.g. images, text)
- Diversity of topic
- Diversity of viewpoint
- Diversity of genre (e.g. blogs, news, comments)
- Diversity of language
- Geographical/spatial diversity
- Temporal diversity

More specifically, dimensions of diversity can be considered at document level and at statement level[3]. With respect to images, the dimensions of diversity listed above can be considered as well. More specifically, the dimensions of diversity for images include:

- Rights owner (person or professional agency who took the picture)
- Time (date and time the picture was taken)
- Location (where it was taken)
- Source (where it was published, e.g., Website, blog, forum, PDF document)
- Production device (if the picture is computer generated, if it comes from a digital camera or a scanner)
- Intent (it is used to attract the attention of observers, to convey emotional messages, to give information for documenting a given claim)
- Sentiment (positive, negative or neutral)
- Context (characteristics of the text surrounding the picture, e.g., background of author, considered aspect, theme of the text)
- Subject (words that describe what the picture shows and that can be linked to the same-similar terms contained in the surrounding text)
- Settings (date and time the picture was taken, as well as broader notions of time such as night/day or summer/winter)
- Style (words describing the style of the photos, e.g. photorealistic, pictorial, etc)
- Pure visual diversity (how visually similar or dissimilar images are, color-wise, structure-wise or both)

Besides the diversity of the Web content itself, the **diversity in queries** might be relevant. Here possible dimensions include the gender of the user, his mood and age, his personal background and the time of query writing.

3.3. Knowledge Evolution

When analysing diversity, opinions, and bias, one often needs a reference point: ***facts*** about politicians, business events, and semantic relationships among people, organizations, and other entities. Such facts are considered to be true, but are subject to a temporal dimension. In other words they evolve in time. For instance, the statement "*Bill Clinton is the president of the United States*" has been true for a certain time in the past, but it is not true anymore. We see time as one fundamental diversity dimension. Issues about their representation and harvesting are addressed later in the report.

[3] These issues are extensively addressed in WP4 (see the report "Bias and Diversity: Approach, Methods, and Algorithms")

4. Technological approach to the solution

Our objective is to enhance the state of the art by developing search facilities that determine diversity in a completely automatic way and capture diversity in all its dimensions. In this section, we describe our proposed solution and how it is mapped with the technologies that will contribute to it.

Fig. 2 shows the basic building blocks and the flow of information of the proposed framework. According to the *user queries* (1) and the specific application (e.g. text or image search, faceted search, future prediction[4]), necessary *metadata* (2) will be extracted and processed (on-line or off-line) from a relevant set of multimodal documents, i.e. the *document corpus* (3), by using a set of suitable *tools* (4). During the whole process, domain specific *background knowledge* (5) will be used to support semantic tools and to fix a common vocabulary both at the level of document annotation (metadata) and queries.

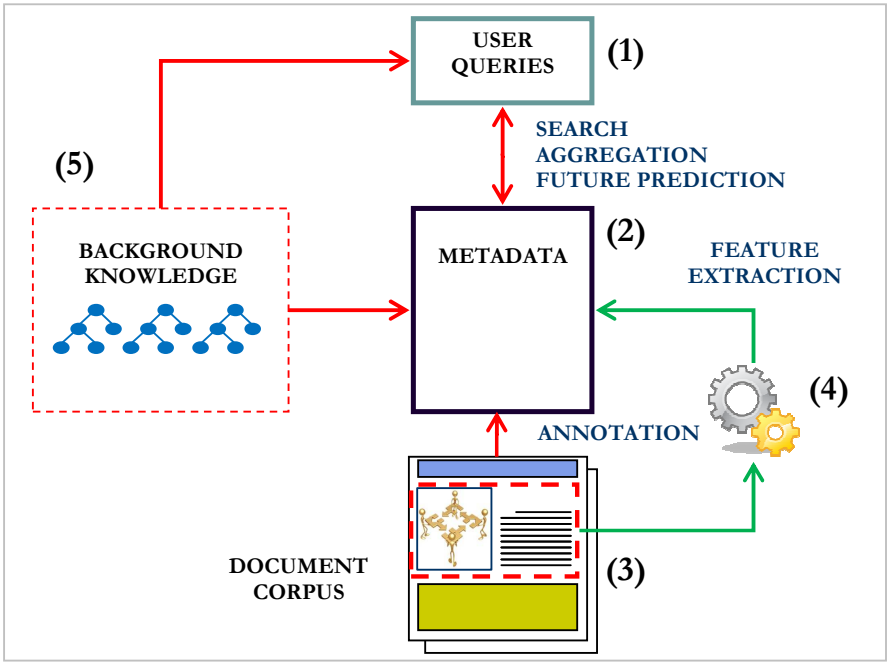

Fig. 2 – The proposed framework

More in detail:

(1) **User queries** express user requests, for instance a textual query or a higher level query constructed by interacting with the user interface of the specific application used. Query results might be a plain list of documents or pictures (e.g. in the classic text or image search) or a complex integrated view summarizing the different diversity dimensions identified (for instance in terms of results aggregated by opinion, main topics, actors, etc. similarly to faceted systems[5]). The main issues here are how to identify and present query results to the user[6];

(2) **Metadata** are extracted from documents and are of fundamental importance to respond to the user queries. They include feature vectors and document indexes. As it may result natural, at this level we need powerful indexing solutions[7];

(3) **The document corpus** is the set of documents which are analysed[8];

[4] Future Predictor will be one of the applications that will be part of the LivingKnowledge testbed (WP6-WP7).

[5] See for instance the Flamenco initiative: http://flamenco.berkeley.edu/

[6] See also the work in the context of WP6 (delivery D6.1).

[7] Again addressed in the context of WP6 (delivery D6.1).

(4) **Feature extraction tools** are the technological solutions responsible to process the document corpus and identify specific parts in them which contain relevant information. They will mainly work off-line for the extraction of the diversity dimensions from documents[9];

(5) **Background Knowledge** encodes domain specific knowledge in terms of concepts and semantic relations between them (see how it could be done in [215]). It is used to give formal semantics to both metadata and user queries. Notice that this is essential to ensure interoperability as the recent trends in the semantic web suggest (see for instance [224]).

For instance, in Section 2 we have provided the following two examples of queries:

- What are the main {concepts, people, parties, countries, dates, resolutions, etc.} related to integration?
- Which of these {concepts, people, parties, countries, dates, resolutions, etc.} are most {controversial, accepted, subjective, biased, etc.}?

Corresponding metadata will include the list of {concepts, people, parties, countries, dates, resolutions, etc.} and opinions extracted from the documents which are relevant to the integration topic. All the {concepts, people, parties, countries, dates, resolutions, etc.} and opinions should be expressed using the available background knowledge to ensure interoperability.

To address these queries, we need tools able to detect and extract the main {concepts, people, parties, countries, dates, resolutions, etc.} from the text, images and possibly other media contained in the documents, detect opinion and bias and process/categorize/aggregate/compare/visualize the results. All such tools should be diversity-aware, namely they should be able to capture diversity dimensions by extracting features, indexing documents and presenting results accordingly. Current state of the art tools are presented in Section 6.

As described in the next section, to process data it will be fundamental to follow appropriate consolidated methodologies (see next section). One of the main roles of technologies will basically be trying to automate, adapt and possibly improve such processes.

[8] In our settings they are maintained by European Archive and crawled by Yahoo! Search Media.

[9] See all the work in WP2 and WP4.

5. Interdisciplinary contributions to the solution

We strongly believe that dealing with diversity requires a high interdisciplinary approach. The following methodologies are considered fundamental contributions to the solution:

- **Media Content Analysis (MCA)** from a social sciences perspective. The analysis typically starts from the formulation of some specific ***research questions***, in terms of topics, actors and patterns of interpretation (i.e. indications about how the discourses are framed) that need to be investigated. The work proceeds with the identification of specific ***variables*** (i.e. the metadata), which make up the ***Codebook***. It consists of different characteristics for every variable to ask specifically about in the relevant media, and of the instructions for the manual coding. The set of relevant media (e.g. documents, newspapers, websites, blogs and forums) is called the ***document corpus*** (equal to a sample in social sciences). In particular, variables are extracted on different levels of the documents: some address the whole document and its source, some focus on claims. Note that the term "claim" is taken from the recently-used method for analyzing public discourse (i.e. political claim analysis) and hence denotes "the expression of a political opinion by physical or verbal action in the public sphere" [23]. We refer to "claim" in a more general sense of "statement" as the expression of an opinion in the public sphere. The variables from the Codebook, which are further aggregated into ***indicators***, are used for statistical purposes when addressing the research questions. The significance of this methodology lies precisely in its capacity to detect context and cultural diversity.
- **Multimodal Genre Analysis (MGA)** from a semiotic perspective. MGA is a two-step process. First text-and-image combinations, i.e. multimodal ***meaning-making units***, are identified and annotated in Websites. Then, they are grouped into a set of ***hierarchical patterns*** (the MGA templates) including, *inter alia*, genres and mini-genres such as logos, contact information, menus, 'running text' paragraphs. Detailed analysis of such patterns, functioning on different scalar levels, helps predicting where specific information will or will not be found in a Website. Inspired by Halliday's theory of meaning, which posits the existence of at least three separate meanings intertwined in every communicative act, this approach views opinion, bias, and other appraisal systems, as part of interpersonal meaning [54][55] and not, in themselves, as part of what Halliday calls ideational meaning, i.e. the expression of ideas. In this view, language and other semiotic resources such as colour, gesture, gaze, shapes, lines etc. are pattern-forming systems which govern the relationship between interpersonal and ideational meaning-making systems. This approach thus has the potential to detect patterns and to predict where to find relevant information and opinions and bias vis-à-vis more factual information.
- **Facet Analysis (FA)** from a knowledge representation and organization perspective. FA is the process necessary for the construction of a ***Faceted Classification (FC)*** of a domain [70][71]. An FC is basically a set of taxonomies, called ***facets***, which encode the knowledge structure of the corresponding domain in terms of the standard terms used, concepts and the semantic relations between them. For each domain, facets are grouped into specific fundamental categories which capture the basic knowledge components in the domain. Originally, Ranganathan [70][71] defined five fundamental categories: Personality, Matter, Energy, Space and Time (synthetically PMEST). Later on, Bhattacharyya [78] proposed a refinement which consists of four main categories, called DEPA: Discipline (D) (what we call a domain), Entity (E), Property (P) and Action (A), plus another special category, called Modifier. For instance, in the medicine domain (D) the body parts (E), the diseases which affect them (P) and the actions taken to cure or prevent them (A) are clearly distinguished. Modifiers are used to sharpen the intention of a concept, e.g. "infectious disease". An FC is typically used to classify books in the domain according to their specific meaning, in contrast with classical enumerative approaches. They have a well-defined structure, based on principles, and tend to encode shared perceptions of a domain among users, thus providing more organised input to semantics-based applications, such as semantic searching and navigation [208]. Our challenge will be to adapt and scale such methodology to the Web.

Fig. 3 shows how these methodologies are integrated into the overall framework. Black boxes correspond to the methodologies described above.

- **MCA** is central. Most of the work in the project aims to automate this part of the process. See the vertical line including points (1)(2)(3). Manual coding is substituted by (semi-)automatic feature extraction and statistical analysis is substituted by indexing and aggregation techniques.
- Conversely, **FA** is horizontal. During the first phase, based on the analysis of typical research questions (which correspond to the user queries), Codebook templates (the metadata and rules to fill them) and the document corpora used in MCA, FA is used to generate the FCs (i.e. the background knowledge) for the domains of interest (look at the three arrows pointing to the Faceted Classifications box). Notice that this process is done off-line and is typically carried out by domain experts. In order to test the framework, we chose as the topic: "European elections: migration, xenophobia, integration" as our use case and identified the following relevant domains: Political Science, Sociology, Psychology, Economics, Law, Geography, History, Philosophy, Religion and Information, Mass Media Research and Communication. Corresponding domain specific FCs have been generated by our Library Science experts. They can be easily extended or adapted later. This is a well known feature of FCs [215]. During the second phase, as described in [208], which extends the work in [209][210], FCs are then used as a controlled vocabulary during the whole process (see the arrows from the Faceted Classification box to the research questions and the Codebook). Background knowledge is used both at the level of tools (4), metadata (2) and user queries (1) to disambiguate and ensure interoperability.
- **MGA** contributes by identifying areas in the documents which are relevant to the extraction of specific information, for instance for opinions and bias. MGA is therefore functional to the set of feature extraction tools, which are meant to automate the annotation processes of the methodologies, are used to fill the Codebook (represented for instance as a set of feature vectors and indexes) with the information extracted from the document corpus.

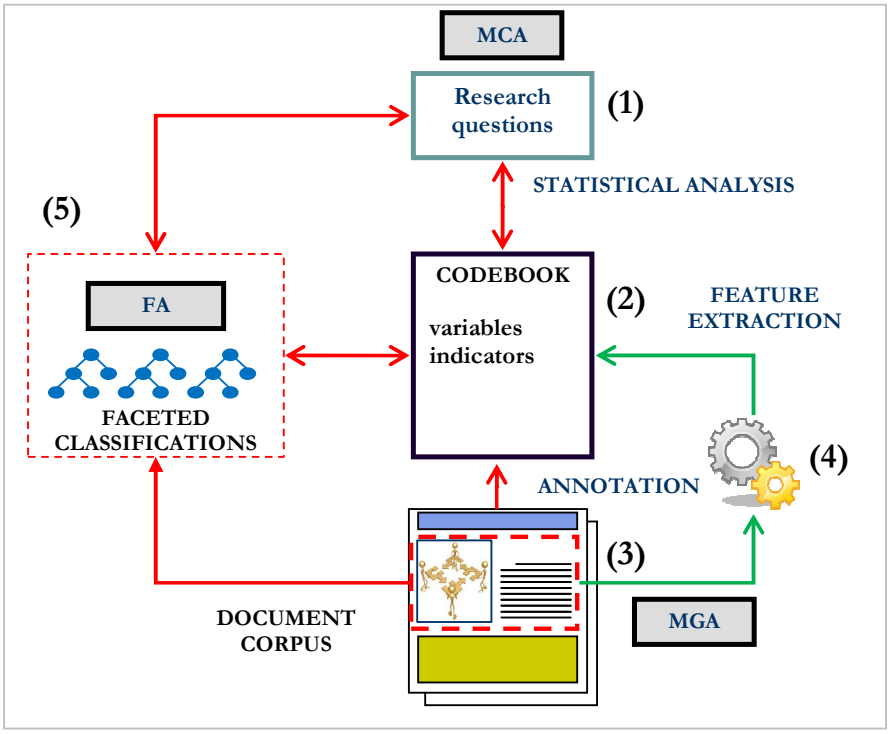

Fig. 3 – Technological integration of the methodologies contributing to the solution

The methodologies are extensively described in the Sections 7, 8 and 9.

6. State of the art

6.1. Generic feature extraction and clustering tools: an introduction

Any automatic analysis of text or images involves an initial stage where basic features such as statements in text and objects or low-level features in images are extracted. Developments in natural language processing that identify and annotate statements include entity tagging, entity resolution and relationship extraction. Of these, entity tagging is an almost solved problem for most common entity types (e.g., person or organisation names) and software tools, e.g. ANNIE/GATE [109] are available. Relationship extraction between entities can be carried out with unsupervised, for instance [102][103][114][125][127][132], or supervised machine learning approaches, e.g. [138].

Facts and opinion extraction

General fact extraction is a far more complex process than relation extraction and should be built upon relation and predicate argument extraction [130].

Extracting opinions from free text is a key task in opinion mining, since all subsequent mining, retrieval, and analysis stages crucially depend on it. Many existing efforts have focused on sentiment analysis that determines semantic orientation (i.e., whether a positive or negative sentiment is conveyed). The basic approach is to look at co-occurrence patterns with paradigm words (e.g., "excellent" and "poor") [112][120][136] or by exploiting additional data such as word glosses [113]. Other approaches detect semantic orientation by applying machine learning techniques [111][137]. For product reviews – a specific kind of opinion of high commercial interest has been studied to produce a concise summary [128][133][107][121]. Additionally, automatic approaches to filter untruthful reviews (that try to manipulate the customer) have been studied [112]. Other methods exploit the "knowledge of the masses", by utilising massive user bases to derive opinions about news [116].

Key notions in our definition of bias are the polarity of an opinion and the context of the opinion holder. Several techniques to identify and analyse opinion are available: opinion mining, or sentiment analysis, has been mainly considered as a binary or three-class classification problem. Applied techniques include natural language processing and machine learning [87] which are mostly applied to online product reviews. Some research explores the problem of identifying the opinion holder; Kim et al. [125] exploit lexical and syntactic information; Kim and Hovy [126] analyse the semantic structure of sentences and use semantic-role labelling to label opinion holder and topic; Bethard et al. [108] propose a semantic parser-based system which identifies opinion propositions and opinion holders. In the latter system, the semantic parser labels sentences with thematic roles (e.g. Agent and Theme) by training statistical classifiers and is endowed with additional lexical and syntactic features to identify propositions and opinion holders. Work on relating opinion holders with their personal background is still unavailable. However, some techniques *do* consider diversity as discussed below.

Diversity in search

Diversity of search results in text retrieval has been considered in the context of result diversification. Since user queries may well be ambiguous as regards their intent, diversification attempts to find the right balance between having more relevant results of the 'correct' intent and having more diverse results in the top positions. In order to improve user satisfaction, the top-k results are either ranked by diversity [104][128] or diversified optionally by clustering them according to the different diversity dimensions covered [115][131].

Diversity in queries is mostly related to user intent when posting a search query. Existing research in this area deals with classification of user queries according to content destination (e.g. informational, navigational, transactional) [126][106]. Some values for diversity dimensions, as considered here, are certainly available through meta-information (e.g. source and resource dimensions). The identification of other values requires automatic algorithms for topic detection, language identification, information extraction and opinion mining.

Image search engines with diversification of search results constitute a relatively new area of research, where one way of increasing diversity is to ensure that duplicate, or near-duplicate images in the retrieved set are hidden from the user [105], e.g. by forming clusters of similar images and showing one representative for each of them. Other approaches use semantic Web technologies to help increase the diversity of the search results. For instance, in ImageCLEF [122] image search results are presented as columns corresponding to the individual topics discovered.

Context analysis identifies relevant information behind the content, especially spatial and temporal information. With images, such techniques can identify the original source of the picture which may be of better processing quality, or even for automatic tagging [135] (e.g. tags propagated from one image to another).

Features for diversity dimensions in images can be extracted directly from EXIF information inserted automatically (digital camera) or manually (e.g. by the photographer) in an image file. In the absence of EXIF tags, some features can be derived using image-retrieval techniques [110], forensic techniques [134][129], and algorithms for automatically annotating images and extracting high-level semantic features [119].

Text Genre, Register, Type

Genre and register are abstractions that characterise the cultural and situational (e.g. spoken vs. written) context of language use. One fairly influential researcher in statistical approaches to genre/register analysis is Douglas Biber. [192] provides an overview on the typical techniques used. Rather shallow, abstract features such as part-of-speech, closed word classes, and type/token ratio are analysed by means of dimensionality-reduction techniques such as factor analysis and cluster analysis to characterise texts on dimensions such as “informational vs. involved” or “overt expression of persuasion”. To avoid confusion with the inherently situational defined notions of genre/register, Biber uses the notion of ‘text type’ to refer to clusters characterised by such automatically extracted dimensions. His most recent book [193] constitutes a more comprehensive overview on these techniques.

A common criticism of Biber’s quantitative, data driven approach to register analysis is that the elicited dimensions and text types are sort of ad hoc. Michael Halliday's work (e.g. [194]) is a good source for a more methodological, linguistic account on genre and register that can be used for qualitative, model-driven register analysis.

It should be noted, however, that bridging the gap between quantitative and qualitative approaches to register analysis is an ongoing scientific and social challenge. Qualitative approaches are based on a rich prior and to some extent generic model that is used to analyse text manually. The inherent mapping between linguistic surface features and model depends much on human interpretation. Quantitative approaches elicit typically flat models automatically, but these models tend to differ from analysis to analysis. [195] constitutes an interesting recent attempt by Biber and Jones to bridge this gap (from an admittedly quantitative perspective). They automatically segment research articles into discourse units, use factor analysis to reduce Biber’s typical surface feature set to four dimensions, and cluster analysis to elicit a text type for each discourse unit.

For LivingKnowledge understanding the reflections of (hidden) situational factors on observable linguistic features is apparently relevant. Immediate applications include clustering search results by genre or visualising mentionings of some entity along time and some situational dimension.

Web Genre Classification

Meyer zu Eissen and Stein [196] investigate genre classification of Web search results. Based on a user study and a pragmatic manual analysis, they distinguish the following (Web) genres: help, article, discussion, shop, portrayal (non-private), portrayal (private, aka personal homepage), link collection, download. They provide an in-depth analysis of useful features for automatic genre classification, focusing on features that can be computed with reasonable effort (lookups and simple POS-tagging). They distinguish among (1) presentation features (such as average number of <p> tags), (2) closed word sets (such as average “word class”, or average number of currency symbols), and (3) text statistics (such as average number of punctuations), and (4) POS (part-of-speech) categories (such as average number of nouns). Reported average classification accuracy (with SVM) on the 8 genre classes tested with 10-fold

cross validation on 1200 documents is about 70%, ranging between 55% and 80% for the individual genres.

Boese and Howe [197] explore Web genre evolution across time. In particular, they are interested in how robustly Web genre can be automatically classified when the training corpus and the test corpus are deliberately drawn from distinct periods (in the range of ~ 6 years). They observe that genre classification is remarkably stable in such periods; however, this may be in part due to their corpus design. They carefully align (fairly small) available corpora used for Web genre classification with their temporal counterparts via URL. Robustness of genre classification may then be partially due to the fixed domain. However, at least they took care that the training set and test set were disjoint in terms of URLs. Reported accuracies for their classification machinery lies at around 75% (55% - 87%) for the Meyer zu Eissen corpus, and at around 80% (50% - 90%) for the WebKB1997[10] corpus crawled from 4 universities (genres: course, department, student, faculty).

This little survey is by no means exhaustive, but gives an overall impression: Web genre classification uses non-topical, stylistic surface features to group Web pages into genres. No agreed upon taxonomy of genres exists and, in all probability none will ever exist. The boundaries between genres are soft, i.e., Web pages may belong to more than one genre. Accuracies for single label genre classification lie at around 75%, some portion of the remaining 25% may be due to the soft boundaries.

Topic Classification and Clustering

Classification (also called Supervised Learning) is the process of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purpose of predicting the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). In the context of document classification our objects are the documents and we aim to automatically assign thematic labels like “Sports”, “Music”, or “Computer Science” to these documents.

> *Formal Problem Setting: let* $\mathbf{d} = (d_1,\ldots,d_M)$ *be the document vector to be classified and* $c_1,\ldots,c_K$ *the possible topics. Further assume that we have a training set consisting of N document vectors* $\mathbf{d_1},\ldots,\mathbf{d_N}$ *with true class labels* $y_1,\ldots,y_N$. N_j *is the number of training documents for which the true class is* c_j.

There exists a great variety of classification methods such as Naive Bayes, kNN, Rocchio, linear SVM, etc., all of which operate on a high-dimensional feature space usually constructed from word occurrence frequencies in documents (and possibly some additional input such as anchor texts in hyperlink neighbours, neighbour topics, etc.). Well known software packages comprising classification algorithms include WEKA [198] and SVMlight [200]. Accuracy depends on the specific categories and number of training documents available and is typically between 70 and 95 percent.

Unlike classification, which analyses class-labelled data objects, clustering analyses data objects without consulting a known class label (unsupervised learning); i.e. class labels are not known and training data is not available. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximising the intracluster similarity and minimising the intercluster similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate *taxonomy formation*, that is, the organization of observations into a hierarchy of classes that group similar subclasses together. In our context, we aim to cluster documents into groups of thematically related documents.

Software packages for clustering include CLUTO [199] and WEKA [198].

[10] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

Frameworks for Information Extraction from Text

Information extraction is concerned with extracting structured information from text. From a fairly abstract perspective this usually involves the following steps:

- ***Segmentation***: chunking text into segments. Typical examples of this are sentence splitting and tokenization.
- ***Classification***: labelling segments with some class. Typical examples of this are POS (part-of-speech) tagging, noun and verb phrase extractors, and named entity labelling.
- ***Normalization***: mapping segments to some equivalence class. Typical examples of this are stemming and lemmatization (at a low level), and co-reference resolution (at a higher level).

Of course concrete extractors often perform more than one of these steps at once.

The past 10 years' technology for information extraction has matured along two main lines. Firstly, for some low level tasks, such as sentence splitting and POS tagging fairly robust components have been developed. Secondly, component-oriented frameworks for information extraction have been developed.

The two most well known frameworks are GATE[11] and UIMA[12]. Both frameworks use so-called annotation graphs (see, e.g. [201]) for imposing possibly multiple layers of structure on text (note that standard XML cannot be directly used for expressing multiple, possibly conflicting layers of structure), and on this basis make it possible to compose information extraction components into processing pipelines. While GATE and UIMA follow a very similar design, and, in fact, UIMA components can be integrated into GATE (and in principle also vice versa), there are a few notable differences, too. GATE's annotations need not be formally specified with a type system, whereas UIMA requires types. This, on the one hand, eases working with GATE; on the other hand, it may be one reason for UIMA to reportedly scale better (the other reason for UIMA's being more scalable is that it supports distribution more readily). GATE comes with a rich set of components for all sorts of basic processing tasks, whereas UIMA is a framework, and the set of publicly available components is rather limited.

In the context of LivingKnowledge, the output of some components for information extraction is needed in order to derive a feature vector on documents, segments, sentences etc. For very simple feature vectors, such as bag-of-words from documents, one does not need such information extraction frameworks. However, as soon as some non-trivial steps are involved, it certainly makes sense to use such frameworks for setting up robust processing pipelines, and extracting features from the annotation layers in a principled manner.

6.2. Available tools for Text Analysis

Natural language processing (NLP) tools offer the key technology for extracting opinions and understanding bias and diversity in text. State of the art NLP tools and techniques are being researched by several partners within the consortium and are contributing to the work of the project. UniTN is developing a range of tools all of which are currently trained and evaluated on English text only, but that can be adapted to other languages for which similar training resources are available.

To pre-process the text, text is separated into sentences and tokens using standard rule-based techniques. After tokenization, morphological processing is applied: POS (parts of speech) tagging and lemmatization. The POS tagger has been trained on the Penn Treebank and uses the corresponding tag set (http://www.computing.dcu.ie/~acahill/tagset.html). The training was carried out using the online Passive–Aggressive algorithm [141] and uses the feature set described by Collins [140] with a linear-time Viterbi decoder. It tags 13,000 words per second on a 2.6 GHz AMD machine and achieves a tagging accuracy of 97.3% on sections 22–24 of the Penn Treebank, the standard test set for English POS taggers.

[11] http://gate.ac.uk/

[12] http://incubator.apache.org/uima/

Dependency Parsing and Predicate-Argument Extraction

To extract syntactic and shallow semantic structure, we employ the LTH syntactic–semantic parser [146]. The following figure shows an example of the type of structure created by the parser.

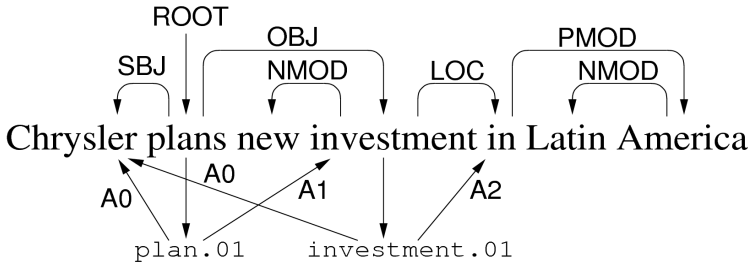

Fig. 4 – An example of structure produced by the syntactic-semantic parser

The outputs of the system are the following:

- Surface-syntactic dependency relations representing the structural relations between words. For instance, *Chrysler* is the subject of the verb *plans*, and *new* is a modifier of *investment*.
- Logical predicates and their corresponding word sense labels (dictionary entry identifiers) in the PropBank [146] and NomBank [145] lexicon. For instance, the verb *plans* corresponds to a logical predicate that is an instance of the PropBank entry *plan.01*. Similarly, *investment* refers to a predicate defined in NomBank as *investment.01*.
- Semantic role relations between predicates and arguments. For instance, the predicate corresponding to the word *plans* has a planner relation to its first argument, the entity referred to by the word *Chrysler*. This is encoded by the label *A0*, which is the identifier of the planner relation in the PropBank entry *plan.01*. Analogously, *Chrysler* also functions as an Investor (*A0*) for the predicate denoted by investment.

The system has a speed/accuracy trade-off option. In the high-speed mode, a cubic-time syntactic parser is used without integration of syntactic and semantic analysis. In this mode, the system processes 450 tokens per second on a 3.2 GHz MacPro machine, while the syntactic accuracy is 85.8 and the semantic F-score 79.5 on section 24 of the Penn Treebank (see [147] for an explanation of these measures). In the high-accuracy mode the syntactic and semantic stages are integrated [144] a syntactic parser with $O(n^4)$ runtime is used. This system processes roughly 7 tokens per second, while the syntactic accuracy is 88.5 and the semantic F-score 81.8. In the high-accuracy setting, the system recently achieved the highest score of all evaluated systems in an evaluation of 20 different systems [147].

The existing system can only handle English text since it was trained on the Penn Treebank corpus of English text. However, since it is a statistical system, it can be re-trained on similar corpora for other languages. Apart from English, there are small hand annotated corpora with syntactic and semantic structures for other languages (including Catalan, Chinese, Czech, German, Japanese, and Spanish [143]).

Subjective Sentence Classification

UniTN has implemented a baseline classifier to separate subjective and objective sentences. It has been trained on 15,753 sentences from the MPQA opinion corpus [148]. The classifier currently achieves a precision for subjective sentences of 0.72 and a recall of 0.75, evaluated via a 5-fold cross-validation over the MPQA corpus.

Currently, the classifier uses a bag-of-words feature representation of the sentence, but plans exist to investigate whether a higher level of accuracy can be reached by using complex linguistic structures. The classifier is a linear support vector machine and was trained using Liblinear implementation [142].

Named Entity Recognition and Coarse Word Sense Disambiguation

Barcelona Media extracts named entities using the Supersense Tagger [139]. This tool outputs three types of tags:

- High-level WordNet sense labels (41 tags)
- CoNLL-2003 named-entity tags (4 tags: location, miscellaneous, organization, and person)
- BBN Pronoun Coreference and Entity Type Corpus named-entity tags (105 tags)

The tagger processes roughly 250 words per second on a 1 GHz AMD machine. The reported performance figures of the word-sense tagger are a precision of 0.77 and a recall of 0.78 on the SemCor corpus [207].

Representing facts

When analysing diversity, opinions, and bias, one often needs a reference point: ***facts*** about politicians, business events, and semantic relationships between people, organizations, and other entities. Ideally, these facts should be systematically organised in a comprehensive ***knowledge base***. Today, there are some sizable knowledge bases like DBpedia[13], Freebase[14], TrueKnowledge[15], or YAGO[16]. They contain millions of entities, along with semantic type information. In particular, YAGO [92] has harvested Wikipedia and integrated the results with the WordNet taxonomy of classes so that all entities in Wikipedia are properly assigned to semantic classes in YAGO.

In the following, we refer to ***facts*** as logical statements that are commonly assumed to be true. They are not necessarily provable, but usually undisputed. Two examples are the following first-order formulas: *shape(earth, sphere)* and *speed(light, 300000km/s)*. A higher-order example is: *happenedBefore(invented(Cerf, Internet), invented(Berners-Lee, Web))*. We focus on ***relational facts*** of the form *predicate(const$_1$, const$_2$, …, const$_m$)* where the arguments *const$_1$* through *const$_m$* of the predicate are individual entities or constant values. Examples are:

politician (BarackObama)
politicalOffice (BarackObama, presidentOfUSA)
meeting (BarackObama, AngelaMerkel, 26-June-2009, Washington/DC).

Without loss of generality, we can concentrate on binary relations. Higher-arity relations can be represented by reifying relational instances and using their identifiers as arguments in other relations. For example, we can represent the meeting between Obama and Merkel as:

id1: meeting(BarackObama, AngelaMerkel)
id2: meetingDate (id1, 26-June-2009)
id3: meetingLocation (id1, Washington/DC)

Our overriding goal is to capture the evolution of this kind of relational knowledge, and to connect it with opinions, their diversity and evolution.

stmt (Plato, believes, shape(earth, disc)),
stmt (Galileo, doubts, shape(earth, disc)).

Knowledge Harvesting

For building and maintaining large knowledge bases, we need to harvest facts about entities and their relations from data sources on the Internet. These sources can be a) structured like databases and factbooks, b) semi-structured like Wikipedia infoboxes and lists, or c) unstructured natural-language text

[13] http://www.dbpedia.org

[14] http://www.freebase.com

[15] http://www.trueknowledge.com

[16] http://www.mpi-inf.mpg.de/yago-naga/

like Wikipedia articles, biographies, news articles, etc. Extracting entities and facts from sources is easy for a), non-trivial but reasonably well understood for b), and difficult for c). The technology for dealing with cases b) and c) is known as ***information extraction (IE)***. State-of-the-art IE methods are (combinations of) rule-based pattern matching, natural language processing (NLP), and statistical machine learning (ML) [89].

With large knowledge bases about typed entities like YAGO or DBpedia, it is fairly easy to detect and extract ***named entities*** (constants in unary predicates) in both semi-structured and natural-language text sources. While NLP-based tools such as GATE/ANNIE[17] or OpenCalais[18] may still play a role, lookups in an explicit knowledge base are often much more effective and efficient. It *does* require entity disambiguation in moving from surface strings such as "Mrs. Merkel" to entities such as *Angela_Merkel*, but this problem is faced by all methods in the same way. YAGO includes a *means* relation between strings and entities, derived from Wikipedia redirections, WordNet synsets, and other sources along these lines. This, together with the type information (e.g., *type(Angela_Merkel, politician)*), provides a great asset when establishing proper mappings.

IE for ***semantic relations*** (binary predicates) between entities is substantially harder and the focus of our work. It can be pursued in an output-oriented targeted ("closed") manner or in an input-oriented generic ("open") manner.

In the case of ***output-oriented targeted IE***, we are guided by a given set of relations for which we would like to gather instances. We are flexible in choosing our sources (e.g., we can opt for easier or cleaner sources with high return) and we can exploit redundancy on the Web. Moreover, we have great flexibility regarding how deep natural-language text is analysed in a demand-driven manner. Snowball [90], Text2Onto [84], LEILA [91], DARE [99], and SOFIE [93] are examples of such IE methods, with Wikipedia being the most prominent source of raw information. A typical use case is to find the *Alma Mater* of as many scientists as possible or many CEO's of companies, using a small set of seed facts for training.

In the case of ***input-oriented generic IE***, we are focusing on a given input source (e.g., a particular Web page, news site, or discussion forum) and consider all conceivable relations at once. This approach inevitably requires deep analysis of natural-language text, and can only be successful if sufficient training data is provided. Here training data typically has the form of fine-grained annotations for complete sentences or entire passages (e.g., the PropBank [86] or FrameNet corpora[19]). This kind of training data is much harder to obtain. TextRunner [82], StatSnowball [101] and tools for semantic role labelling [96][85] fall into this category of IE approaches. A typical use case is to automatically annotate news and extract as many relations as possible from each news item.

Input-oriented generic IE is more ambitious than output-oriented targeted IE, but requires many more computational efforts for natural-language analysis and critically depends on the availability of sufficiently large and representative training data. Thus, output-oriented targeted IE usually achieves significantly higher accuracy, and is generally more robust. In the LivingKnowledge project, we will initially pursue output-oriented targeted IE methods. We are compiling a catalogue of selected key relations, along with specific seed facts, that are particularly relevant for IE and opinion analysis about politicians, business entrepreneurs, actors, and singers. Note that targeted IE of this style yields not only new facts but also extraction patterns. These patterns can, in turn, be used for input-oriented IE on arbitrary sources such as news, as long as we focus on the same set of interesting relations.

MPII applies SOFIE [151] in order to extract facts, especially about individuals from unstructured texts to extend the YAGO ontology [150]. The definition of fact is based on that of statement. A statement is a tuple of a relation and a pair of entities. A statement has an associated truth value of 1 or 0, which is denoted in brackets. A statement with truth value 1 is called a fact. For example,

bornIn(AlbertEinstein, Ulm) [1]

[17] http://gate.ac.uk/ie/annie.html

[18] http://www.opencalais.com

[19] http://framenet.icsi.berkeley.edu/

where *bornIn* is the relation between entity *AlbertEinstein* and *Ulm*. So given a document *d* and a relation *r*, SOFIE is capable of extracting a list of individuals of *r* from the text.

In addition, the LEILA [1] system extracts a list of term pairs for a given relation *r* without the need of an underlying ontology. For the relation *bornIn*, this results e.g. in

bornIn(Albert Einstein, Ulm) [0.8]

from the sentence "Albert Einstein was born in Ulm." where *0.8* is a confidence value.

In order to obtain a high accuracy, the raw texts should be categorised according to domains. A domain should contain a main topic like global warming or several highly related subtopics like various digital camera products.

Opinion Extraction

For the opinion extraction task there should be gold standards for training and evaluation. The gold standards for training and evaluation are manually annotated subjective expressions, opinion source and opinion target in each document. To save the annotation overhead, the annotations are only in domain related sentences.

- ***Subjective expression***: a subjective expression is any word or phrase used to express an opinion, emotion, evaluation, stance, arguing etc. In [148], it is referred to as a private state. The text span of a subjective expression is a minimal constituent covering a private state
- ***Opinion source***: the person or entity expressing the subjective expression, possibly the writer. Each opinion source is a noun phrase
- ***Opinion target***: targets of the subjective expressions. More specifically, the concept, entity, or fact about which the opinion is expressed. Mostly a target is a noun phrase, if a target is expressed in terms of a clause containing a fact, the target can be represented as a fact

Each subjective expression should contain an attribute indicating the polarity of the opinion, if it expresses a negative, neutral or positive opinion. It is also desirable to annotate the opinion intensity (low, medium, high, extreme) of the opinion expressions if the time is allowed. In addition, if there are several targets or opinion sources associated with different subjective expressions, the corresponding associations should be annotated.

For the sentence "We fear an early death much more", "fear" is the subjective expression, "we" is the opinion source and "early death" is the target. The polarity is negative and the intensity is high.

Besides the expression-level annotation, a document should be assigned with a global opinion, if the writer holds a strong negative, weak negative, neutral, weak positive, strong positive opinion about the main topic. The document level annotation is also important because many documents contain a mixture of positive and negative opinions. Hence, it is desirable to identify the dominated ones.

Simple opinions only use *likes/dislikes* modalities about simple statements such as individual entities. Likes/dislikes are often referred to as ***polarities***, and their strengths may be additionally quantified, leading to the following form of simple opinions: *opinion (subject, modality, strength, statement or object)* [87][101]. Examples are:

opinion (user1, likes, 0.9, BarackObama)
opinion (user2, likes, 0.7, AlGore)
opinion (user2, dislikes, -0.3, politicalOffice(AlGore, VicePresidentOfUSA))

6.3. Available tools for Image Analysis

With the advent of digital media the role of images in the communication process has steadily gained significance. In particular, a single image can hardly convey precise information or detailed data on a given event/subject, but an image can often transmit, in a more effective and immediate way, a message or an emotion. The use of visual data encapsulated in textual data powerfully enriches the

communication process that the writer is performing. Another consideration concerns the role of photography as a means for the faithful and true reproduction of real events, and the use of images when documenting facts to be described. Two aspects need to be considered: on one hand, photographers by taking pictures choose their own way for reporting an event (as do writers); on the other hand, pictures may be manipulated before their use, thus conveying different information with respect to their original intent. Hence, the value of photography as a record of events must be established carefully. To summarise, images have three main roles within a communication process, i.e. they can be used:

a. **To attract the attention of observers**: a picture can be included in a document to attract attention and make the document more appealing. See for instance Fig. 5
b. **To convey emotional messages**: images used to document a real situation, but at the same time conveying an emotional message with a negative implication. See for instance the picture in Fig. 6; the messages of drought and flood reach the readers in a very immediate way; in Fig. 7 the image reporting a polar bear is used, in contrast to the one in Fig. 5, to convey a negative message, e.g., the ice melting as a consequence of global warning
c. **To convey information when documenting a given claim**: images can be used to reproduce and support a claim, e.g., the existence of global warming (e.g. Fig. 7).

A polar bear in its natural sea ice habitat. (NOAA Photo Library)

Fig. 5 - Example of a picture included in a document for attracting the attention of observers.

Considering these important roles of images, it is clear that technological means for image analysis would be extremely useful for the evaluation, understanding and validation of pictures used in the communication process.

A ship on the river Rhine in Duesseldorf, Germany, July 25, 2003, during the extreme heat wave that scorched Europe for much of the summer. Low water levels meant bigger ships could transport only 30 to 50 percent of their normal cargo. (AP Photo/Martin Meissner)

Devastation in coastal Venezuela caused by the December 1999 flows of mud and rocks. (Lawson Smith, U.S. Army Corps of Engineers)

Fig. 6 - Example of images that convey an emotional message with a negative implication.

Fig. 7 - An image used to convey an emotional message with a negative implication, in contrast with Fig. 5.

Left: In 1978, the Qori Kalis Glacier looked like this, flowing out from the Quelccaya Ice Cap in the Peruvian Andes Mountains. Right: In 2002, the view of Qori Kalis has changed dramatically with a massive 10-acre lake forming at the ice margin. (Courtesy of Professor Lonnie G. Thompson, Byrd Polar Research Center, The Ohio State University)

Fig. 8 - Example of images used to document a given claim.

It has been recognised from the outset that multimedia aspects of documents would in general play a supporting role in the extraction of facts, opinions, bias and diversity rather than a central role. Probably the most direct contribution will come from CNIT's and UNITN's work on Image Forensics and the detection of image manipulation that, in some cases, may provide evidence that a biased view is being projected.

Content Based Image Retrieval

Content Based Image Retrieval (CBIR) systems aim to organise and find images in large databases based on their low-level features. The retrieval process usually relies on presenting a visual query to the system, and extracting the set of images that best fit the user request (query-by-example mechanism) from an image database. This process involves the extraction of a set of descriptive features that are used to perform the matching between the query and the target images according to some distance measures. Several years of research in this field [152][153][154] have highlighted a number of problems related to this process. In particular, images with high feature similarities to the query image can be very different from the query in terms of the interpretation made by a user (user semantics). This is referred to as the semantic gap, which characterises the difference between low-level image content and high-level semantics. Several additional mechanisms have been introduced to achieve better performance. Among them, Relevance Feedback (RF) is a very powerful tool to iteratively collect information from the user and transform it into a semantic bias in the retrieval process [155].

UniTN has developed a content-based image retrieval system for low-level features extraction from the entire image or from regions such as: color histogram, moments of the color distribution, edge histogram and wavelet texture histogram (see MPEG-7 standard [156]). To reduce the semantic gap, UniTN uses two different feedback mechanisms: human feedback or automatic feedback based on pre-classification of image corpora. Relevance feedback allows the system to collect information from the user and transform it into a semantic bias in the retrieval process. The system uses relevance feedback and a stochastic algorithm called Particle Swarm Optimizer (PSO) [157] in tandem. PSO is a population-based stochastic technique that allows complex optimization problems [158] to be solved. In recent years, PSO has been used as a way to generate optimised parameters for several algorithms and has also been proposed also in the field of CBIR. In [159] a detailed study of the use of statistical methods in image retrieval problems was recently presented, where the image retrieval task is treated as a classification problem. A very preliminary version of UniTN's work was presented in [160], where the concept of PSO-CBIR was first introduced. In this system, retrieval is formulated as an optimization process which is iterated until convergence. Image ranking in the form of an ordered list is provided as the final product (various interfaces are also allowed). In order to evaluate CBIR systems, performance evaluation measures have been proposed based on their precision (number of relevant images retrieved/total number of images retrieved) and the recall (number of relevant images retrieved/total number of relevant images).

In the future, UniTN plans a more complete low-level feature library both organising all the feature vectors in a hierarchical way and exploiting metadata information. They also intend to contribute to building a corpus of data, linking documents with concepts in an ontology or taxonomy. Finally, using concept networks, they plan to link up similar entities, objects, people, according to visual features and semantic concepts.

In order to carry out these tasks, UniTN requires images with certain characteristics, i.e. photos or computer generated images (no tables or graphs or any kind of line drawings). However, for their future work, to carry on the research activity on opinions and emotions conveyed by images, they will need information about opinions linked to images and perceptual features that characterize such opinions (semiotic features, e.g., colour or perspective information). Then, to improve the performances of their CBIR system, they will need contextual links between images, and annotations about concepts, people, places and events on a set of images, used for training their algorithm.

Image Forensics Technologies

In today's digital age, the ease of creation and distribution of digital images causes their importance to increase day by day. But a co-occurring disadvantage is the ease with which digital content can be manipulated, casting increasing doubt on its validity as an accurate and trustworthy representation of reality. Digital images are used in the communication process to convey a given message or an emotion in an effective and immediate way, or as a means for the faithful and true reproduction of real events, to document facts and support a given claim. But the availability of low-cost hardware and software, useful for modifying digital images in their aspect or semantics, imposes the parallel advent of technologies needed for assuring that what we are seeing in a photo is a true representation of what really happened. Recently, image forensics has been largely proposed as a valid technological means for ensuring the credibility of digital images, by both extracting knowledge about the origin of the content [161] and detecting the application of a wide variety of manipulations [162]. Image forensics is based on the idea that inherent traces (like digital fingerprints) are left behind in digital media during both the creation phase and any other subsequent processes [163]. By relying only on the analysed data, without any previously embedded information (passive approach) and without the knowledge of the related original data (blind method), forensic techniques capture a set of intrinsic information carried out on the digital asset by means of different analysis methods (e.g. statistical, geometric, etc.) and provide useful information on image history.

Knowledge about manipulation suffered by images may provide evidence that a biased view is being projected. The exploitation of technological solutions can help in the analysis of pictures, by making important details more obvious that human observers can barely perceive or report during manual image annotation. By looking, for instance, at the first photo in Fig. 9 there is no evidence of

the tampering that has occurred, but by applying forensic algorithms (in particular those developed in WP4 for bias analysis) the manipulated region can be located, as shown in the photo (Fig 5. b). It is the human observer who understands what objects are to be considered as relevant in an image, and thus if interesting objects are specified by manual image annotation, forensic technologies could only be applied to such regions.

a) Tampered image b) Analysed image

Fig. 9 - Tampered and analysed image

In the context of forensics, CNIT is carrying out research activities in both main directions: i.e. source identification and tampering detection, in order to acquire a set of features characterising images. In fact, a forensic algorithm can be seen as a forensic block taking an image as input and providing some information as output which make it possible to better evaluate, understand and validate pictures used in the communication process.

Regarding information on content origin, the aim of a forensic block is to identify the source that produced the picture. Specifically, CNIT are working on:

- a forensic block that can determine whether the picture is computer generated or from a camera [164][165]
- a forensic block that can determine whether the picture comes from a digital camera or a scanner [166]

Usually, such methods are based on a classifier, thus requiring a large image corpus for training the classifier before fine tuning the system.

Regarding the detection of the application of a wide variety of manipulations, different forensic blocks are able to distinguish whether the image suffered different processing operations. They focus on some forensic blocks providing information on:

- ***re-sampling operation***: when geometric transformations are applied (e.g. rotation, scaling) a re-sampling of the original image to a new sampling grid is included [167][168][169]
- ***double JPEG compression***: when creating a digital forgery, it is always necessary to resave the modified image: since JPEG is a very common format, the tampered image may incur double JPEG compression [170][171][172][173]
- ***copy-move forgery***: a part of the image is copied and pasted onto another part of the same image [174][175][176][177]
- ***brightness, contrast and colour enhancement operations***: in order to enhance the image's aspect some operations which adjust brightness [178], contrast [178][179] and colour are applied to images

Usually such methods work for good-quality (low-compression) images and dimensions of at least about 300 x 300. Accepted formats are RAW, TIFF, JPEG, PNG, but not GIF.

Almost all the mentioned forensic blocks are under development at a more or less advanced stage. The

study and development of some algorithms has still to start (in the next months). However, all blocks will be refined during the project.

UniTN has developed a method that differentiates between photorealistic and photographic images. In particular, they have implemented a forensic technique for automatically detecting local forgeries, i.e., objects generated with computer graphics software that are inserted into natural images, and vice-versa. One of the first approaches in this direction was presented in [180] where wavelet-like decomposition is used for constructing a statistical model for photographic images. Another wavelet-based approach has been proposed in [181]. Also a geometry-based approach has been proposed in [182]. Other techniques exploit imaging sensor pattern noise statistics [183][184]. Recently, a novel approach was also presented in [185] where differences in perception of computer-generated and natural images are analysed.

UniTN has also developed a forensic technique to analyse a common form of manipulation based on the 'composition' of two or more people in a single image, i.e., it allows one to detect composites of people. The approach estimates a camera's principal point from the image of a person's eyes. Inconsistencies in the principal point can be used as evidence of tampering. In fact, in authentic images, the principal point is near the centre of the image. When a person is shifted in the image as part of the creation of a composite, the principal point is moved proportionally. Differences in the estimated principal point across the image can thus be used as evidence of tampering [186].

Other Image Analysis Tools and Techniques

SOTON has a number of tools for analysing image content and context. In particular, they have a collection of tools that generate various low-level image feature vectors. Given an image, these tools are capable of creating a fingerprint for the image. Typically, these fingerprints can be compared or matched to find similar images, or instances of the same image. In the past, SOTON has also been particularly interested in developing feature vectors that allow searches for a particular object, or sub-image, within a large collection of images. These feature vectors may also work when the query image exhibits large amounts of geometric distortion with respect to the images being matched [189].

At a higher level, they have tools that allow analysis of image content at a level that humans can comprehend. For example, tools are available [191] to determine:

- whether or not an image is in focus
- the number and positions of the faces in the image
- whether the depicted scene is indoor or outdoor
- whether the depicted scene is urban or rural
- whether a flash was used etc.

SOTON is also working on "automatic-annotation" systems, e.g. [187][188]. These systems apply machine learning techniques, which given enough training data (i.e. pairs of <feature vectors, list of words describing an image>) are able to learn how to associate low-level image features with words describing an image's content in a human language. Once the automatic-annotator has been trained, it can be used to predict annotations or words for new, previously un-annotated images.

It should be noted that the performance of all of these techniques is highly dependent on many factors [187]. In particular, the choice of feature vector morphologies and the amount of training data can have a massive impact on how well an automatic annotation system works. Additionally, automatic annotation systems do not work equally for all words. For example, they might be able to learn a good representation of what a Dalmatian looks like, but not what a Terrier looks like.

Finally, SOTON has developed software that is able to semantically link images with contextual information. For example, given a photo with time/location information, they can determine what the weather was like when the photo was taken, what stories were in the news at that time and what the nearby geo-tagged Wikipedia articles are.

For input, SOTON software basically requires images in suitable formats (such as JPEG or PNG). They can transform image formats automatically if required. Training classifiers and automatic annotators sometimes requires substantial amounts of annotated image data. Annotations for this could, for example, be provided manually, or extracted automatically from collateral text within the document

embedding the image. Keywords, captions, and “alt text” are a good source of data for this, and it is also possible that natural language processing techniques could be employed to find relevant words/phrases/sentences/paragraphs in the document itself. Some ideas for new functionality to be developed within LivingKnowledge include:

- The extension of SOTON’s own and other existing robust image matching strategies [189][190] to provide annotations and facts. For example, they would like to investigate automatically determining geo-location of the image by matching against large sets of geo-tagged images, determining the original source (i.e. the particular image in a photo-agency’s catalogue) of the image, and propagating knowledge from visually similar images of the same subject. Such knowledge may be obtained from multiple sources, specifically DBpedia and GeoNames, but also for example from collections of data accumulated specifically for an LK scenario. The Southampton team has been accumulating economic information, weather and climate information and geographical information. Common properties such as time and location may be used to connect data together, so for example a query such as 'was it sunny when this image was taken' is possible by using metadata and/or image matching techniques to determine when and where the image was taken, and then cross-referencing this with the collected weather data.
- Investigating and developing improved auto-annotation models. In particular, they are interested in scalability and active learning (i.e. instead of training once at the beginning, the system is always learning new relations). They would also like to investigate the use of better annotations, and the fusion of multiple-image feature types.
- Extending and developing novel techniques for the dynamic incorporation of available contextual information. In addition, the ability to be able to fuse the outputs from both contextual analysis with content analysis in order to build a much deeper understanding of an image. As an example, assume someone runs a query of a photo of the Eiffel Tower. The image matching algorithm might reason that it is either Blackpool Tower or the Eiffel Tower but not be able to say which with certainty. If the training set can be marked up with basic factual knowledge that the Eiffel Tower is in France and Blackpool Tower is in the UK, then a better estimate as to the identity of the structure in the query photo can be made.

6.4. Temporal knowledge

Temporal IE extends fact extraction (see section 6.2) by determining the time point or time interval for which a fact is valid [100][88]. This is usually determined from the same sentence or passage from which the fact itself is extracted. But multiple extractions for the same fact (from different sources) can be combined to strengthen temporal information. This can be in the form of refining the temporal resolution (e.g., the exact date, rather than only the year reference, for the start of someone’s tenure of a political position), filling incomplete information (e.g., unknown end of the term for a political position), or invalidating false hypotheses by consistency checks (e.g., positions seemingly held after a person’s death).

Temporal expressions can be ***explicit expressions*** like dates (e.g., “July 15, 2009” or “July 2009”) or ***implicit expressions*** like adverbial phrases (e.g., “until the end of this month”, “last week”, “years later”). The former can be extracted by regular expression matching, the latter require deep natural-language analysis (e.g., dependency parsing) and/or a good dictionary of temporal expressions [94]. For both it is often necessary to a) validate that they actually refer to the considered fact (and not to another aspect of the same sentence) and b) determine the exact denotation that connects the fact and the temporal expression. For example, an expression may denote the beginning of an interval during which the fact holds true, its end, both, or a relative time-point or interval. These steps also need some form of natural-language analysis, ranging from part-of-speech tagging for very simple sentences to dependency parsing for complex sentences.

In addition to these temporal aspects of relational facts, another issue to be studied is the evolution of knowledge as a whole. This could manifest itself in changes of underlying taxonomic structures or terminology. For example, what changes has the Wikipedia category system undergone in the last decade? Which are due to the growth and better structuring of Wikipedia and which reflect the evolution

of the Zeitgeist and corresponding knowledge structures? Similar questions can be asked about terminology evolution [95][83].

Fact-Media-Opinion Co-Evolution

Facts about entities (people, movies, companies, political parties, etc.) evolve over time. New facts are added (e.g., awards, lawsuits, divorces), some facts change (e.g., spouses, CEOs, political positions). These changes do in turn influence the media coverage of certain entities. Facts and media coverage together influence opinions in specific portals, blogs, forums, etc. This situation is illustrated in Fig. 10. Potentially, there are also influences from media and opinions on facts. For example, media coverage may force a politician to resign from an office, and sometimes it is the grassroots' opinions in blogs and online forums that are eventually picked up by print media and TV. Understanding these mutual influences, as a function of time, is a key issue in the LivingKnowledge project.

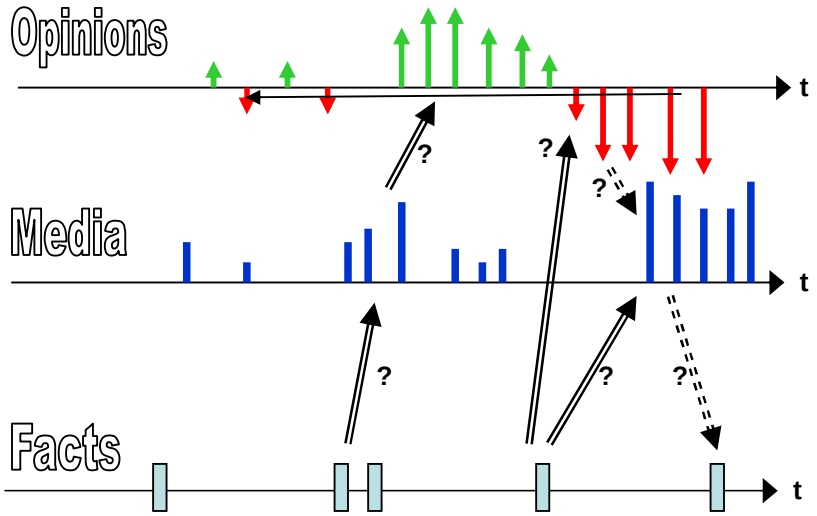

Fig. 10 - Illustration of Fact-Media-Opinion Co-Evolution

Ideally, we could view this problem as a joint probability distribution of entities, facts, media measures, opinions, and time (with corresponding attributes added, e.g., intensity of media coverage, polarity of opinions, etc.), but this seems to be infeasible in its full extent. Thus, we plan to address sub-issues of reduced complexity first, and later aim to generalise.

Capturing and analysing media coverage and opinions in online communities entails identifying suitable sources and extracting relevant attributes from them. The relevant attributes include the entity of interest, the source itself, the date (and sometimes the time), the opinion holder, associated numeric data (e.g., ratings), associated textual data (e.g., reviews and other comments), and whatever additional attributes are available about the opinion holder (e.g., home city or country, gender, age, etc.). Some of these extractions are easier than others: striving for a "sweet spot" of high return for low effort is crucial. In particular, it is often difficult to compile sufficient training data for learning a robust model and ground-truth data for systematic evaluation. In this regard, corpora that include numerical ratings are particularly valuable, as these can be a "soft substitute" for ground truth. Examples are movie portals like www.imdb.com or www.rottentomatoes.com, which contain a large number of user reviews and ratings over an extended time period, or discussion forums on prominent people in entertainment, business, and politics such as www.thewiplist.com, which contain temporal profiles about people whose assessment in the public has been highly time-variant such as Bill Clinton, Silvio Berlusconi, etc. As for the evolution of media coverage, Websites like news.google.com provide relevant information, as shown in Fig. 11. In isolation, this kind of information is not that informative. But when connected with temporal facts and opinions, it can reveal important insights. For example, the April 2008 burst in Berlusconi's timeline and the older burst in 1994 are clearly due to his becoming elected as prime minister. So, facts and media intensity are highly correlated. Similarly, we could predict a new media burst about Bill Clinton because of his trip to North Korea in early August 2009.

The co-evolution of facts, media coverage, and opinions requires two kinds of models: a) an explanation model to identify the key factors that determine certain changes in media coverage or opinions, and b) a prediction model to forecast trends and shifts in opinions given recent observations about new facts and media coverage. Mathematically, these could be the very same model, solved in backward and forward direction. Constructing such a model needs ingredients from time-series predictors and categorical classifiers. For example, a time-aware logistic regression model for relations [97] could be a promising direction.

Fig. 11 - Media intensity timelines (source: news.google.com)

7. Contributions to the solution from the social and political sciences

In this section we present Media Content Analysis (MCA) from a socio-political perspective. In particular, we provide an example for the integration phenomenon use case. The first paragraph provides a list of relevant research questions. In the second paragraph, we briefly describe the methods and levels of MCA for the identification of variables and indicators (from a relevant set of media) which constitute the Codebook. Collected data are statistically analysed to respond to the research questions.

7.1. Methodological background

Content analysis can be described generally as "systematic reading of a body of texts, images, and symbolic matter" ([38]:3). It "is applied to a wide variety of printed matter, such as textbooks, comic strips, speeches, and print advertising" ([38]:6) or more generally to any cultural artefact.

As a common definition for further work we suggest the following: Content Analysis is an empirical method for (I) systematic and inter-subjective understandable **description** of textual and formal characteristics (II) and for inquiring into social reality that consists of **inferring** features of a non-manifest **context** from features of a manifest written text and other meaningful matters [40][39][38].

Applying this kind of Content Analysis, we are able to describe content from different sources. We can ask for contingence of certain characteristics within texts. We can also ask for interpretations that are used by people who are cited in the media. "Accordingly, content analysis of texts afford answers to questions about 'what themes occur', 'what semantic relations exist among the occurring themes', and 'what network positions are occupied by such themes of theme relations' among texts with particular types of source, message, channel, or audience." ([41]: 2701)[20].

7.2. Relevant research questions

Based on the theoretical discussion above, we have settled on the following descriptive research questions as the most relevant:

- What ***topics*** occur to what extent in the mediated discourse on integration?
- What ***actors*** in what roles are present in the mediated discourse on integration?
- What ***patterns of interpretation*** (frames) occur in the mediated discourse on integration?

Integration here refers to the two chosen sub-topics (labour market and religion). Within these questions different analytical questions can be raised, for example:

- What are the main [*topics, actors and countries, arguments, frames*] related to integration?
- Which of these [*topics, actors and countries, arguments, frames*] are the most [*controversial, accepted, subjective, biased, etc.*]?
- Who are the main [*politicians, parties, organisations etc.*] discussing integration in a [*negative, positive, neutral*] context?
- Which [*politicians, parties, organisations etc.*] have changed their discourse on integration (i.e. from positive to negative)?
- What time periods are most important for integration, and what other events are correlated to these periods?
- How developed is the discursive character of statements made by different [*politicians, parties, organisations etc.*]?

In terms of input-output analysis (official material and press releases vs. mass media content, forums and blogs):

[20] A more detailed description of the methodology of media content analysis in the social sciences is given in WP8.1. report.

- To what extent have the certain political parties' ideas been reported in the media, blogs and forums of official sites?
- What differences exist between input (communication on Green Websites, press conferences) vis-à-vis output (discussion in media, blogs and forums)?

7.3. Levels of analysis

7.3.1. Coding units

In order to identify and characterise the different opinions contributing to the selected immigration issues, the analysis process starts from a relevant set of media (e.g. Web pages) about the topic. The media content is coded (annotated) at three different levels (a.k.a. coding units):

- ***The whole article.*** We look for publication type, topic, date, reason for publication etc.
- ***The statements in the article.*** We look for speaker(s), their affiliation, content of the statement etc.
- ***The patterns of interpretation*** within one statement. We ask for frames of interpretations for different topics.

Notice that one article can contain one or more statements. In mass media studies, it is common to analyse four statements per article, which is normally sufficient for journalistic articles. However, this restriction arises mainly because the analysis process is manual. The articles and related statements are analysed with a quantitative, structured content analysis. The patterns of interpretation (frames) are analysed with a qualitative, inductive content analysis.

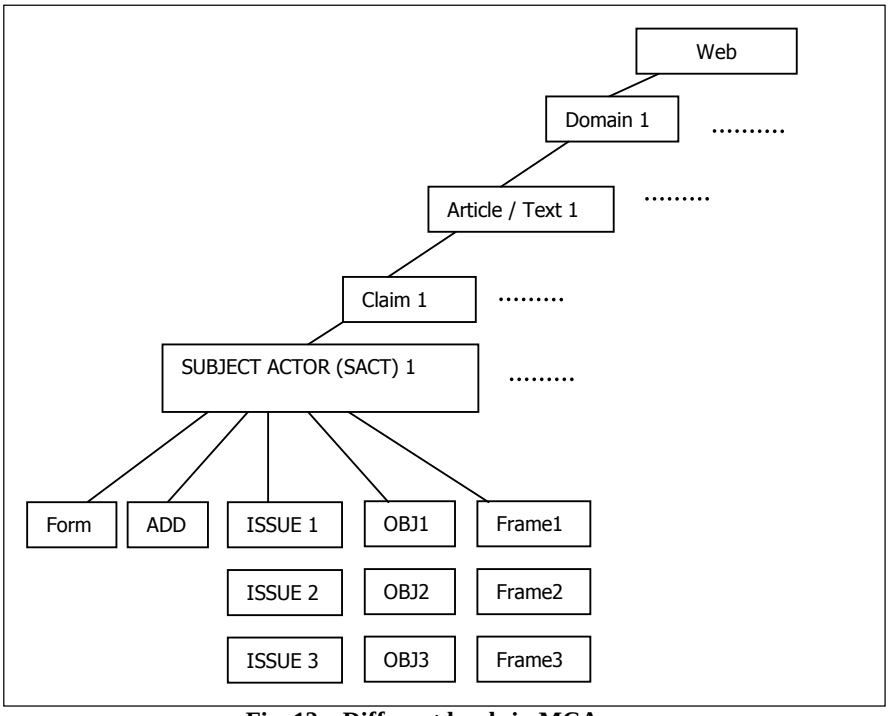

Fig. 12 – Different levels in MCA

7.3.2. Variables and indicators

Based on the main research questions, we can identify different indicators (a.k.a. categories) at different levels for MCA. Each indicator consists of different variables, which ultimately have to be operationalised for empirical analysis. Variables and indicators make up the Codebook (an example is given in Appendix A).

.....

Indicators at the level of article are:

- ***Descriptive variables*** for the description of media content
- ***Topics*** in terms of cognitive-cultural integration, structural integration, social integration, and/or emotional integration
- ***Actors*** (active and passive speakers, objective actors), which refers to "speaker" and "addressee" on the statement-level
- ***Argumentation***, which refers to "issue" on the statement-level, in terms of:
    - (1) Assimilation, (2) Integration, (3) Segregation, segmentation, separation (4) Marginalization
    - Positive, neutral, negative opinion

Indicators at the level of statement are [24][23][33]:

- ***Location*** of the statement in time and space (when and where was the statement made?)
- ***Speaker***, who is the actor articulating the statement (who produced the statement?)
- ***Form*** of the statement (e.g. political action, verbal statement etc.)
- The ***addressee*** of the statement (to whom is the statement directed?)
- The ***substantive issue*** of the statement (what is the statement about?)
- The ***object actor***, namely who is affected by the statement (for/against whom?)
- The ***justification*** for the statement (why should this action be undertaken?)

Following Koopmans ([23]: 3), who applied this to claim analysis, a statement can be transformed into grammatical terms in the following way: "a SUBJECT-ACTION-ADDRESSEE-ACTION-OBJECT-JUSTIFICATION CLAUSE sequence: an actor, the subject, undertakes some sort of action in the public sphere to get another actor, the addressee, to do or leave something affecting the interests of a third actor, the object, and provides a justification for why this should be done.".

Notice that the term "claim" is taken from the recently used method for analysing public discourse (i.e. political claim analysis, e.g. [23]). Claim is used with a direct linkage to political statements and is defined as "the expression of a political opinion by physical or verbal action in the public sphere." [23]. As public discourse in general might not only consist of political statements, we refer to "claim" in a more general sense of "statement" and define it as the expression of a fact, a claim or an opinion by physical or verbal action in the public sphere. To ensure the connection to academic discourse, we will use the term "claim" as well.

Indicators at frame level include possible patterns of interpretation (topics) of statements. Possible subtopics for frames have to be identified. This has to be done on a theoretical basis, which means that former studies, empirical results, and hypotheses have to be used for the creation of the subtopic-list. In our use case (European integration process), we know from different empirical studies that identity frames, instrumental frames, and historical frames are of special interest. See [33] for an extensive description. For instance, an identity frame that occurs in a statement could be "What is (or should be) the EU (not); what does (or should) it (not) stand for?" ([1]: 61). To deduce the frame-indicators for the Codebook, the part of the media content and the list of different arguments, which refer to both frames need to be parsed. Based on this list, the whole content is analysed.

7.4. Issues relevant for the chosen topic

In this section, we briefly introduce the topic "European elections: migration, xenophobia, integration" from a socio-political perspective.

7.4.1. Integration of Migrants – a European Topic

Integration and migration are debated in the academic and political context in a lively way. This entails a few preliminary remarks for the further discussion of the use case:

(1) **Terminology.** The terminology, which is used in these debates, is not always congruent. Different notations are used for the same phenomena and vice versa. We opt for labels deduced from the state of the art in social science debates and the most common terms used in European discussions.

(2) **Sub-areas.** The analysis of integration requires the identification of sub-areas within this complex range of topics. These sub-areas are of analytical need for the content analysis. We will pick up some sub-areas for the in-depth analysis of mediated content and we will use some analytical distinctions for operationalisation and creation of categories. These selections and analytical distinctions should not be understood as shortening the complex field of integration studies.

(3) **The role of media.** In modern media societies, mediated information and discussions about integration and migration are part of the integration process so that the analysis of media content is highly relevant when attempting to understand integration.

Integration (of Migrants)[21] has become an intensely, sometimes fiercely, debated issue in Europe in recent decades:

- The Internal Market comprises the free movement of persons (and citizenship), including free movement of workers, and freedom of establishment
- Migration within the European Union (EU) and from outside the EU has created increasingly diverse societies with considerable numbers of residents who are foreign-born or have foreign-born parents
- The role of Islam has been debated for various reasons, e.g. in the context of the process of Turkey's integration into the EU
- Right-wing populist parties in many European countries have campaigned against new migration, immigrant groups, as well as against European integration.

For demographic reasons, however, the EU and the EU member states will have to attract immigrants in the years to come, and this competition "calls for policy co-ordination and for sustained efforts in the area of integration to ensure equal opportunities for the actors involved" [31]. As diversity will, very probably, increase further and the question of distribution of wealth between, and within, societies will predominate, integration will probably have great relevance for Europe in the coming decades. It is also safe to assume, that during election campaigns like those for the European Parliament, controversial issues like migration and integration will be brought to the centre of public attention. One party, that has always highlighted migration and integration policies, are the European Greens. As they are pioneers in running a coordinated campaign (in the sense of cross-national coordination) during European Elections (see [7], [26]), they are the natural choice for analysing a party's input in the field of migration.

In the academic, as well as in the political, debate about integration, we recognise a diversity of opinions about integration:

- Is integration a process of assimilation, or is integration also possible when immigrant groups maintain certain cultural differences [4]?
- Is integration more an effort by the receiving society or should immigrants pursue their own integration?
- Will it be necessary to attract more immigrants in the future, or are there already enough residents from third-party countries in Europe?
- Should immigrants be selected by qualifications or not?
- Can a growing Muslim population in Europe create mostly solvable challenges, or are Islamic and European culture incompatible?

[21] Integration is used here in the sense of integration of migrants in the context of their receiving society, not in the sense of Economic or Political Integration increasing the supranational collaboration of the EU member states.

- Should integration (see [20]) take place before social integration in order to enhance it or should legal integration (like naturalization) follow a process of social integration?

7.4.2. Dimensions of Integration

The goal of the study is to reconstruct the mediated discourse on integration, to describe the relevant actors and patterns of interpretation and to explain differences between actors, topics and different degrees of diversity. The basic research questions are:

- What (sub-)topics occur to what extent in the mediated discourse on integration?
- What actors in what roles are present in the mediated discourse on integration?
- What patterns of interpretation occur in the mediated discourse on integration?

Classic approaches to integration and migration studies were based on the assumption, that assimilation of immigrants into the receiving society will take place in due course of time. This approach was based on studies in the US in the 19th and early 20th century. This approach is still relevant in the academic discussion on integration, with various improvements: different types of integration like acculturation [3] and social integration [8] are discussed.

Four degrees of integration are ideally distinguished (see [18]: 200f):

(1) Assimilation, which means completely accepting the norms of the majority and forsaking one's original identity

(2) Integration, which means maintaining one's original identity while simultaneously integrating into, and accepting the norms of, the majority

(3) Segregation, segmentation, separation, which means maintaining one's original identity and isolating oneself (partially) from the majority

(4) Marginalization, which means the loss of one's identification with one's origins but failure to identify with the majority, which leads to isolation

Following Asser ([8]: 289; [9]: 27), integration can be differentiated into different areas:

- Cognitive-cultural integration, which pertains to knowledge and skills
- Structural integration, this refers to the question as to whether and what economic or social positions are held by migrants
- Social integration, which refers to relations between people
- Emotional integration, which refers to identification with the incorporating society, acceptance of morals and values

A slightly different distinction, which refers more to the macro level of society, can be made vis-à-vis ***"system"*** and ***"social" integration***. With reference to David Lockwood, system integration is understood as "the orderly or conflictful relationships between parts". Social integration is understood as "the orderly or conflictful relationships between the actors" (Lockwood, in [8]: 268). System integration refers to integration in institutions, organizations, the economic market and, especially relevant for media societies, integration into mediated and the public discourse [15]. Social integration is directly linked to actors and their orientations, motivations, relations and intents.

From this distinction, we separate two main themes for further analysis of mediated discourse: ***we will focus on integration and the labour market as part of system integration. Secondly, we will focus on integration and religion as part of social integration.***

Within these subtopics we will analyse in particular statements and argumentation concerning xenophobia. Xenophobia is one dimension of *Group Focused Enmity* (see [37][44]), like racism, anti-Semitism, homophobia or sexism. Xenophobia can consist of different fears [32]:

- loss of economic capital
- loss of social capital
- loss of cultural capital
- loss of physical capital

These levels will be used to operationalise the relevant characteristics in the written text. Xenophobia needs to be ***distinguished from stereotypes*** (see [36]) attributed to certain groups.

7.4.3. The role of mediated information and mass media

Modern societies are characterised as complex and sophisticated. Public communication is mainly mediated. There are different forums for the public sphere (see [22], [6]: 106ff):

- ***Simple public spheres*** (encounter public sphere) with the presence of the communicating people and direct relations between the communicating people
- ***Middle public spheres*** (thematic or assembly public sphere) with the presence of the communicating people but without direct relations between all the communicating people
- ***Complex public spheres*** (mass media public sphere), where many people participate but only a few people are active communicators

The number of forums ranges from simple to complex public spheres, while the power and authority for speaker decisions increases from simple to complex public spheres.

Diverse communication forums have different influences on integration processes. In particular, mass media public spheres are seen as relevant for integration processes on the level of society. Mass media are seen as important agencies of socialization. They affect norms, values, behaviour patterns, mentalities etc. Primary agencies of socialization such as the family and people directly associated are of diminishing relevance and are replaced by mediated information. In mass media societies, mass media content, in particular, has an impact on socialization [25]. Consequently, mass media also have a significant influence on the evolution of collective and individual identities (see e.g. [29]) and are of great interest for understanding integration processes. Simple public spheres are more relevant on an individual level and are relevant especially for interpersonal relations and connected integration processes.

Mass media content can be provided through different technical media: printed newspapers, TV channels, radio stations as well as Websites. Simple and middle public spheres can also be mass mediated, a forum on a Website is a form of middle public sphere, and a chat room is a special form of simple public sphere. To cover the different forms of public spheres, ***we selected Websites from mass media as well as Websites with blogs (a special form of simple public sphere) and forums (middle public sphere) for analysis***.

An important criterion when distinguishing between these forms of content is the author's role. Recent studies distinguish organised, journalistically working actors from non-organised, publishing actors. This has implications for the analysis of content, insofar as journalistic routines have impact on diversity within journalistic texts.

However, different theories of the public sphere emphasise diverse demands on public discourse. Relevant criteria, which will be used for further analysis, are: who participates, in what way, in what role, with what goals [11]. The different theories of public sphere also imply different levels of rationality of the discourse. We use the most demanding model, namely the discursive public sphere [16], as a heuristic model for operationalising and ask for rationality, balance, number of facts (according to Brantner et al. [5]), number of cited sources in a statement ([15]: 98).

Consequently, the research focus is on what the mediated public discourse concerning integration looks like. By analysing written text and photographs in written texts, we try to address the question of who says what to whom with what interest and in what channel (following Lasswell, see McQuail/Windahl 1981: 10 [26]). Within these questions the diversity of speakers with all their characteristics (who), addresses and object actors with all their characteristics (whom) is important, too. Further on, different topics and the diversity of opinions is of high relevance. The LivingKnowledge project tries to grasp these different kinds of diversity.

8. Contributions to the solution from semiotics

In this section, we present multimodal genre analysis of Websites from a semiotic perspective. The analysis is essentially performed in two main steps. First, a set of relevant Websites are annotated identifying *meaning-making units*, namely a set of higher-order units − e.g. text *and* images − in the Web page conveying a specific meaning. Second, a set of *hierarchical patterns* (a sort of syntactic structure) is identified. The presentation is articulated into two main paragraphs. The first may be seen as a characterisation of the objectives of multimodal analysis (with examples relevant for the chosen use case) with the specific goal of identifying opinions and bias, while the second provides specific details about macro-strategies and their identification in Websites. The contribution is viewed against a background of interface design. It respects and builds on the idea, that permeates and underlies the LivingKnowledge Project, that rethinking the relationships between Website designers and users will potentially lead to significant breakthroughs in human-computer interaction. Increasingly semiotics is being called upon to illuminate the partnership between designers and users in the overall communication process that takes place through an interface of words, graphics, and behaviour. For example, the standpoint adopted in [50] is that designers must tell users what they mean by the artifact they have created, and users must understand and respond to what they are being told. By coupling semiotic theory and engineering, this approach to human computer interaction design encompasses affordances for producing meaningful interactive computer system discourse and, potentially, achieves a broader perspective than might otherwise be possible. In the belief that semiotic engineering is an important way forward, the approach adopted below is that there is, first and foremost, a need to understand the possible meaning-making (i.e. semiotic) processes that take place in a Website. The view taken thus goes beyond traditional semiotics which, mostly for historical reasons, focuses on individual meaning-making resources (language or visual) and takes the combined deployment of visual, spatial and linguistic resources as its starting point.

8.1. A multimodal semiotics standpoint for migration in Websites

Given the goals of the project and the chosen use case, this section corroborates, extends and integrates the relationship between the use case, migration, and the expression of bias and opinions. However, the main purpose is to demonstrate that, as described throughout this Section, meaning-meaning processes in Websites are hierarchical (see *Scalar models of genres* in *Appendix D*) and multimodal rather than sequential, linear and linguistic and hence *paradigmatic* rather than *syntagmatic* in nature. The insight that language-based studies cannot accurately represent all the textual/compositional structure of Internet Websites suggests the need to characterise Websites, and everything that is expressed in them, within a whole-page view of Website annotation. In other words, the problem is to understand what Websites are from a meaning-making standpoint. The solution to this problem is provided in 5 steps by establishing that:

- Opinions and bias are part of the interpersonal meaning system rather than ideational or textual/compositional meaning systems;
- Websites rely on integrated visual, spatial and linguistic resources rather than on language in the enactment of all meanings;
- Websites are based on hierarchical rather than linear principles of meaning-making;
- Hierarchical processes and patterns of integration between visual, linguistic and spatial can be reconstructed over a large number of Websites thanks to the use of *MCA Web Browser*[22] *(see Appendix C)* a manual annotation tool with some semi-automatic features;
- Typical meaning-making patterns in Websites can be recovered through corpus-based annotation and through relational database search facilities which are currently being integrated into the *MCA Web Browser* system.

[22] *http:mcaweb.unipv.it*

8.1.1. Opinions and bias are part of the interpersonal meaning system

What is bias? What is opinion? These are non-trivial questions to answer since the meaning of these terms is far from fixed. For example, there are two meanings associated with the term 'bias', one related to prejudice, the other simply to a quantitatively measurable predominant interest. The Collins Cobuild Dictionary illustrates this distinction as follows:

> *Bias is prejudice against one group and favouritism towards another, which may badly affect someone's judgement of a situation or issue. Bias against women permeates every level of the judicial system ... There were fierce attacks on the BBC.*
> *Bias is a concern with or interest in one thing more than others. The Department has a strong bias towards neuroscience.*

The very fact that bias has two almost diametrically opposite meanings one of which is associated with a negative connotation, the other with a more neutral and often positive connotation, provides us with a clue that bias and opinion belong to a system of meaning which is different from that of ideas. That is, these two meanings of bias are indicative of the existence of a system of meaning through which we are able to appraise and judge others' ideas within synchronic and diachronic timescales. While bias and opinion express ideas in themselves, they are present in all discourses (and not just Websites) in such a way as to represent *perspectives* on others' ideas. If not so, it would be impossible to express an opinion, or to show bias. Opinion and bias must thus be construed instead as part of a meaning system which is different from a meaning system concerned with ideas.

To this end, Halliday's theory of meaning (or to use his term 'metafunctions') is a useful starting point since it clearly posits that at least three separate meaning systems are intertwined in every communicative act that we make. Specifically, *opinion, bias and other appraisal systems (see White, Martin) are part of interpersonal meaning* and not in themselves part of what Halliday calls ideational meaning, i.e. the expression of ideas. In this view, language and other semiotic resources such as colour, gesture, gaze, shapes, lines and so on, are systems which function in such a way as to manage the relationship between interpersonal and ideational meaning-making systems. They are linking systems. However, the textual/compositional forms which we use to link up interpersonal and ideational meaning-making systems are themselves meaningful. They are a source of meaning, which Halliday calls 'textual' meaning (but which some co-workers, in an effort to clarify that meaning forms go beyond written or spoken uses of language, describe as 'compositional'). Part of their meaning derives from the fact that all texts (i.e. units of meaning and not just written texts) adopt expected patterns as regards the way interpersonal and ideational meanings are created. Imagine how a young child would react if a bedtime story were not told using expected textual structures (which include for example collocations such as "Once upon a time" and "They all lived happily ever after"). Other examples include the fact that *all* languages have developed textual resources that foreground either ideational or interpersonal meanings as circumstances dictate. The textual systems which English has developed to foreground interpersonal meanings include, for example, tag questions (themselves derived from two other systems: the modal verb system and the question system). Thus structures such as "You will take me on holiday to the Bahamas next year, won't you?" are heavily marked discourse strategies designed to impose the speaker's opinions and will on a discourse partner. They are also indicative of the fact that opinions and bias may often be equated to the subcategory of interpersonal meaning concerned with power relationships in discourse whether at a personal level (e.g. in relation to saving face) or at a community level (e.g. in relation to ideological aspects of meaning making). Website analyses thus need to find new ways to describe and detect the interpersonal/ideational aspects of forums, chats and blogs.

Summarising, semiotic studies of language are related to functions in discourse rather than to grammatical forms *per se*. Thus, in his study of ideational meaning, Halliday [49][50] presents a strong argument for describing English in terms of *Participant-Process-Circumstance* relationships in which Participants are typically expressed through nouns and pronouns, Processes are typically expressed through verbs and Circumstances typically through adverbial and prepositional structures. Particularly significant in this respect is Halliday's characterisation of the English language in terms of six basic verbal processes of which three predominate, namely material, relational and mental processes. How

striking, for example, is the clear division made in English between material processes which are expressed with *-ing* forms and mental processes which are not. Thus, functional approaches to language have led to changes in the way annotation systems work in the light of the realisation that only searching through immediate constituents (i.e. the sequence of individual words) in an effort to discover opinions is likely to bear little fruit. All this suggests that the starting point for automatic opinion detection systems for Websites needs to be based on different premises.

8.1.2. Websites as multimodal and hierarchical meaning-making systems

Within the overall project goals, we use Websites on the topic of (human) migration to demonstrate that Web pages are primarily multimodal objects. They make meanings hierarchically following topological principles relating to the organisation of space, and hence – periodicity over linearity predominating – structures. The way that evolution over time is expressed in the home page in Fig. 13 expresses how multimodal and hierarchical meaning-making principles dominate in contemporary Web pages. Take, for example, the *masthead-cum-logo cluster* used to express the Website's topic in this and many other Websites. These cluster type functions, rather like a title, give the page its basic identity. The *masthead* is the name of the page, i.e. in this case "*Moving Here*", while the *logo* is, in this case, the rightward pointing curved arrow. Additionally, the *masthead-cum-logo cluster* in this, as in many institutional Websites, includes a subtext: "*200 years of immigration in England"*. This text hypotactically (i.e. in terms of a relation of subordination and inequality) extends the meaning of the *masthead* and, in this case, establishes the Website's diachronic dimension and concern with changing opinions over time. Even more striking is this page's use (like thousands of others in the Web) of *hierarchy*, rather than *linearity*, to organise the *page* as a meaning-making unit. This is a demonstration of the fact that textual dimensions of meaning relate to *all* formal resources existing on a Web page in a holistic way and not just to *written text* or – even more limiting – to *running text* i.e. the form of written text that is typically made up of paragraphs. In addition, this page does not contain paragraphs and sentences of the type typically associated with paragraph-based *running text* which typically contains explicit subject and verb nucleus (a.k.a. the rhetorical theme-rheme structure of *paragraphs*). Instead, this Web page contains highly elliptical linguistic structures integrated with visual and spatial resources. The only sentences used are in the imperative form (*Find out... Go to... Contact us...*), which by definition have no subject. Most of these imperative structures are not prioritised but are, instead, thematically and hierarchically subordinated to other structures on the page. Thus the invitation to *"Explore photos, recordings and documents, research your family, history and even add the story of how you came to England"* is in much smaller font compared with the *masthead-cum-logo* cluster which functions as its "owner". In other words, those sentences which do exist on the page (all of the imperative type) are being explicitly demoted vis-à-vis other meaning structures, all indicative of the multimodal and hierarchical nature of Websites. The use of this type of device may have pre-existed the Internet but what is striking is its high and constantly increasing incidence.

We may further illustrate the principle of thematic expansion which is at work on this page. As illustrated by the red arrows in Fig. 14 a reading of the page starts with the Macro-Theme which is expanded into 4 Macro-News thanks to the cohesive tie created by the Swoosh-type object identifiable with the *logo*. To put the matter in a different way, the *logo* functions on a par with numbering systems, tabulating systems, and subordinating structures in language to create ordered expansions of the underlying thematics, thus, creating an index − albeit one that transcends the forms and functions associated with indices in traditional literacy. The *logo* is used to structure and articulate the entire *page*. A different colour is used to indicate that there are four different *Clusters* which make up a higher level unit of meaning based on a repeating visual pattern and hence form a *SuperCluster*.

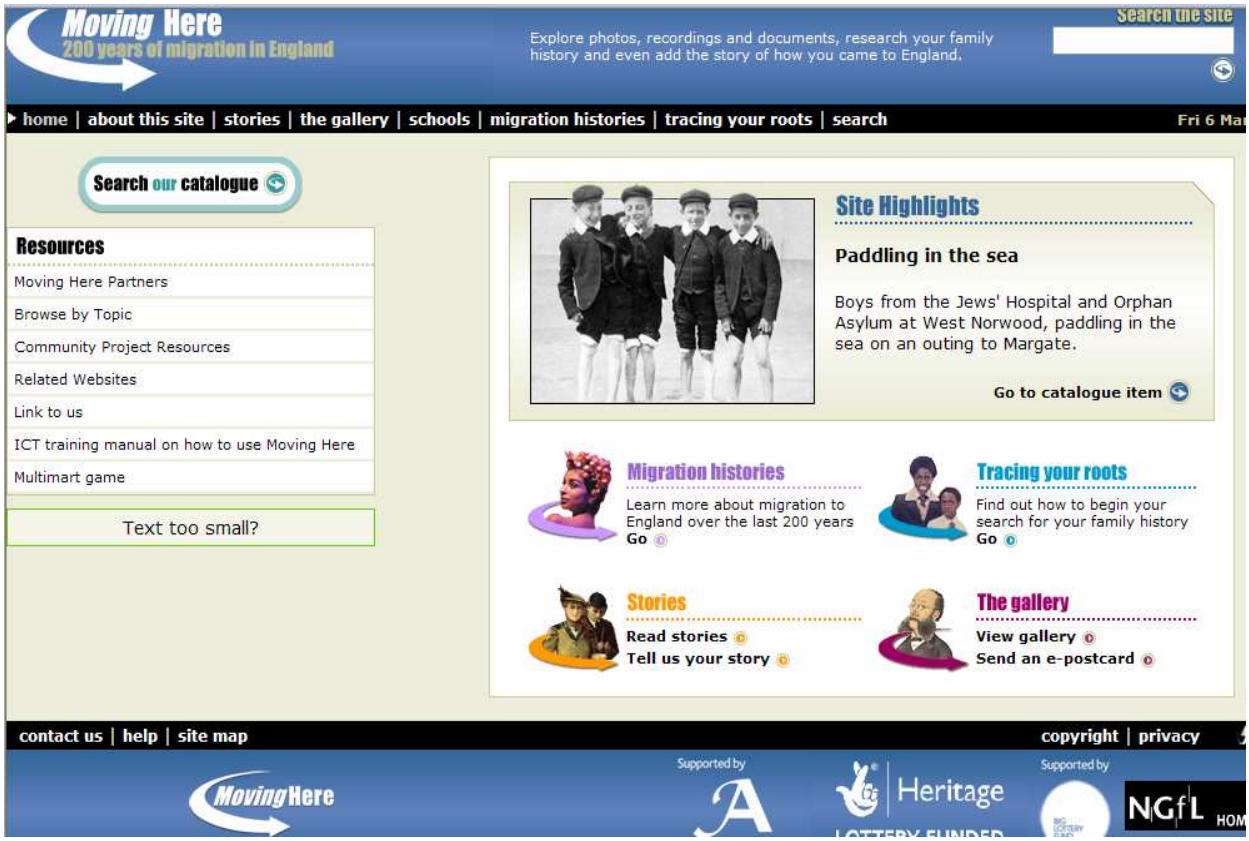

Fig. 13 - The Home Page of the "Moving Here: 200 years of migration in England" Website

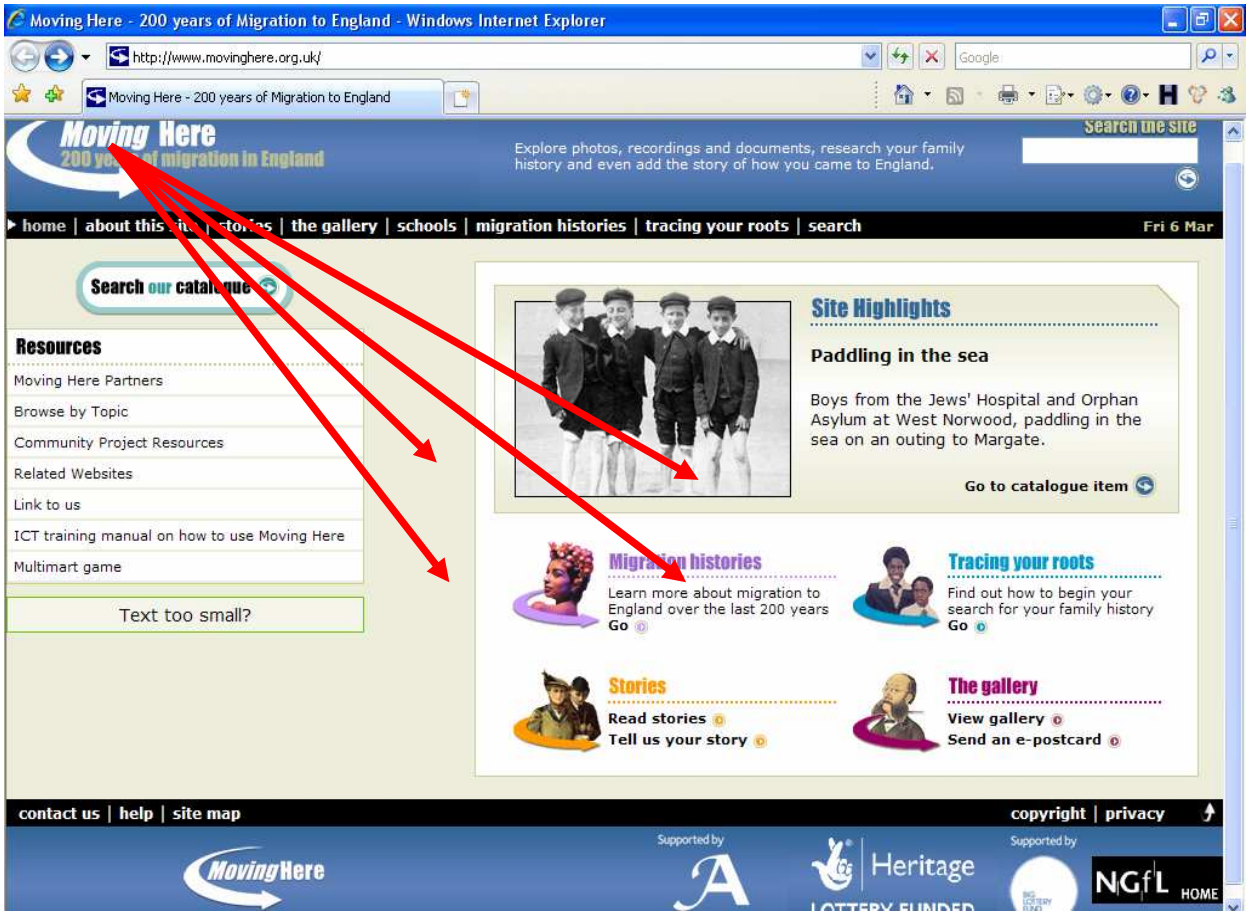

Fig. 14 - Multimodal thematic expansion in the Moving Here Website

When clicking on *Migration Histories* the page shown in Fig. 16 is loaded. The function of the page is to provide an *Introduction* to personal experiences of migration to England over a period of 200 years. With the invitation *"Listen to people's personal experiences of the different receptions they faced when arriving in England, and the struggle to create a new home"* the page sets up the expectation that many different opinions are contained in the Website, many of which are likely to relate to poverty and social relations of power. Indeed, the *Moving Here Web*site is characterised by the prominent use of a *Pagelet*, a higher order textual structure, typically related to strongly ideological stances. Fig. 17 (see dotted rectangle) contains a *Pagelet* timeline which dominates the *page* and functions as a visually oriented timeline complex. Thus, it replaces the earlier generation of timelines with linear paragraph structure with one based on the principle of periodicity. Fig. 15 exemplifies a more traditional timeline consisting of a date followed by a one-line synthesis of an event.

Pagelets are made up of *SuperClusters* which, in turn, are made up of *Clusters* which, again, are made up of *SubClusters.* In the example given in Fig. 16 the *Pagelet* is made up of 4 *SuperClusters* i.e. a set of *Clusters* that contains a periodically repeating pattern such that invariants are easily distinguishable from variants. The first *SuperCluster* is the overall timeline made up of three *Clusters*, i.e. the numbers indicating centuries, which are linked up with each other at the *SuperCluster* level by a light yellow background. The second *SuperCluster* is the *Central Bar* of 7 photos, 4 relating to people and 3 relating to written documents which are again linked up at *SuperCluster* level by a top and bottom light yellow *Bar*. The third *SuperCluster* is the set of 7 *Clusters* each made up of a *Callout* link to numbers indicating a specific year and which contains 124 words, of which only 11 are finite verbs, all further evidence of the strategy which eliminates paragraph-based written text in this contemporary Website. The final *SuperCluster* is the *Caption-cum-Heading* Caribbean Migration Histories Timeline in which the *Cluster* and *SuperCluster* levels are conflated (i.e. are not distinguished). This textual structure consists of 4 words, 1 a noun functioning as the *Thematic Head* (Timeline), two nouns functioning as adjectives (Migration, Histories) and one adjective. It contains no verbs.

We argue that Timelines are:

- multimodal objects which, in today's Internet world, are increasingly replacing evolutionary accounts based on running text as they respond better to the meaning-compression principle and are more intuitively graspable by users
- a typical example of the rise of multimodal Web objects at the expense of language-only Web objects which are disappearing in Web pages, so much so that not all Timelines are explicitly identified by the wording "Timeline", all part of the process whereby Websites are, from a semiotic standpoint, but not perhaps from an information technology standpoint, typically becoming more screeny than pagey
- a typical example of the process of eliminating clause-based wordings (e.g. sentences and paragraphs) in favour of verbless and conjunctionless.

Timelines, like the one illustrated have distinctive linguistic, visual and spatial properties, including the tell-tale, textually-discrete references to years and the explicit line-based linkage between image and written text. Suitably annotated, these features, possibly coupled with keyword searching, would make *Timelines* such as the one shown in Fig. 17 detectable as GENRES and would come to the rescue of scores of teenagers desperately trying to revise history ahead of tomorrow's classroom test who currently waste precious time when resorting to keyword methods that unearth many irrelevant instances of the word 'timeline' and, frustratingly, fail to detect the higher-level *Timeline* genres they are seeking.

A child migration timeline

1607 First permanent English settlement at Jamestown, Virginia, on the north American coast.
1615 Labour shortage. The Privy Council sanctions the transportation of convicted felons to Virginia and the West Indian island colonies such as Bermuda.
1617 London Common Council and the Virginia Company consider sending 'vagrant' children (street kids) to Virginia.
1619 First 100 vagrant children rounded up and despatched to Virginia; venture declared a success; second group planned.
1620 (January) Opposition to child migration; first group sent illegally, but 31 January the Privy Council authorises child migration. Second 100 children sent to America.
1622 Indian massacre of 350 settlers in Virginia in the wake of which another 100 vagrant children were sent among the reinforcements.
New England arranges the emigration of some 200 poor children (not, however 'vagrants') to North American colonies. The children were escorted to the colonies by members of th
ork in Americas had grown to meet the perennial labour shortage in the colonies. Bristol the main port of emigration. Parliament passes an ordinance to make spiriting a felony.
1664 An office was created, under Roger Whitely, to register intended emigrants leaving British ports for the Americas. Spiriting continues; registering not a success.
1682 Privy Council creates another department to register young persons leaving for the colonies to counter the activities of unscrupulous emigration agents and 'spirits'.
1698 (September) Flying Post newspaper reporter claims he observed 'about 200' kidnapped boys held on a ship in the Thames awaiting departure for the colonies.
1740s Aberdeen (Scotland) and environs saw some 500 young people kidnapped for the colonies.
1756 Marine Society founded by Jonas Hanway and Sir John Fielding to train 'boys from 12 to 16' found roaming the streets for service in the Royal Navy.
Peter Williamson exposes kidnapping of children in Scotland for service in the Americas. The book led to a celebrated civil action against certain Aberdeen businessmen and magistrate
can War of Independence confirms the Declaration of Independence of the Thirteen seaboard colonies in North America; end of adult convict transportation to the American colonies
1788 New South Wales established as a penal colony; many of the convicts transported were under eighteen years of age.

Fig. 15 – An example of a traditional timeline

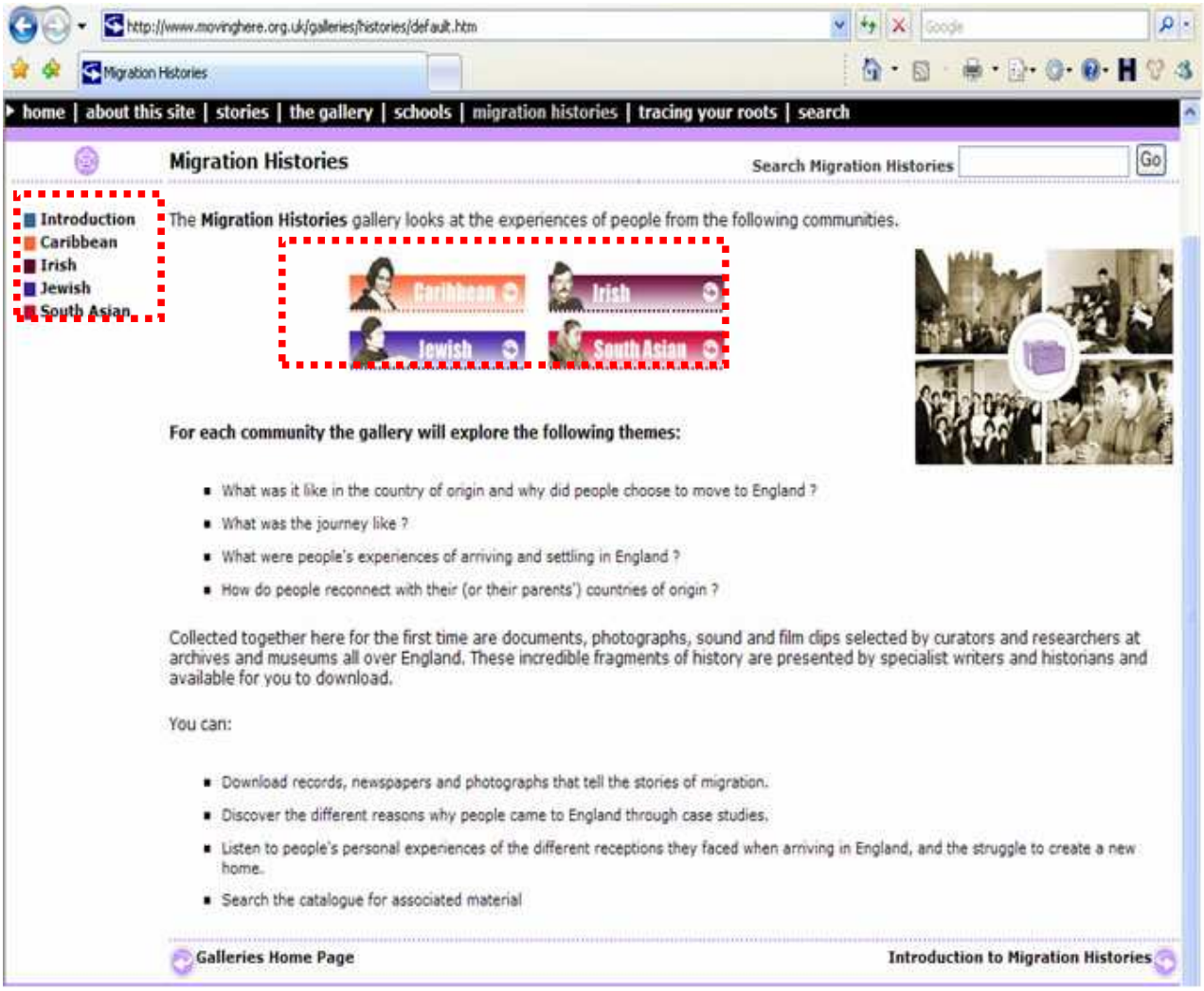

Fig. 16 – Navigating the Moving Here Website

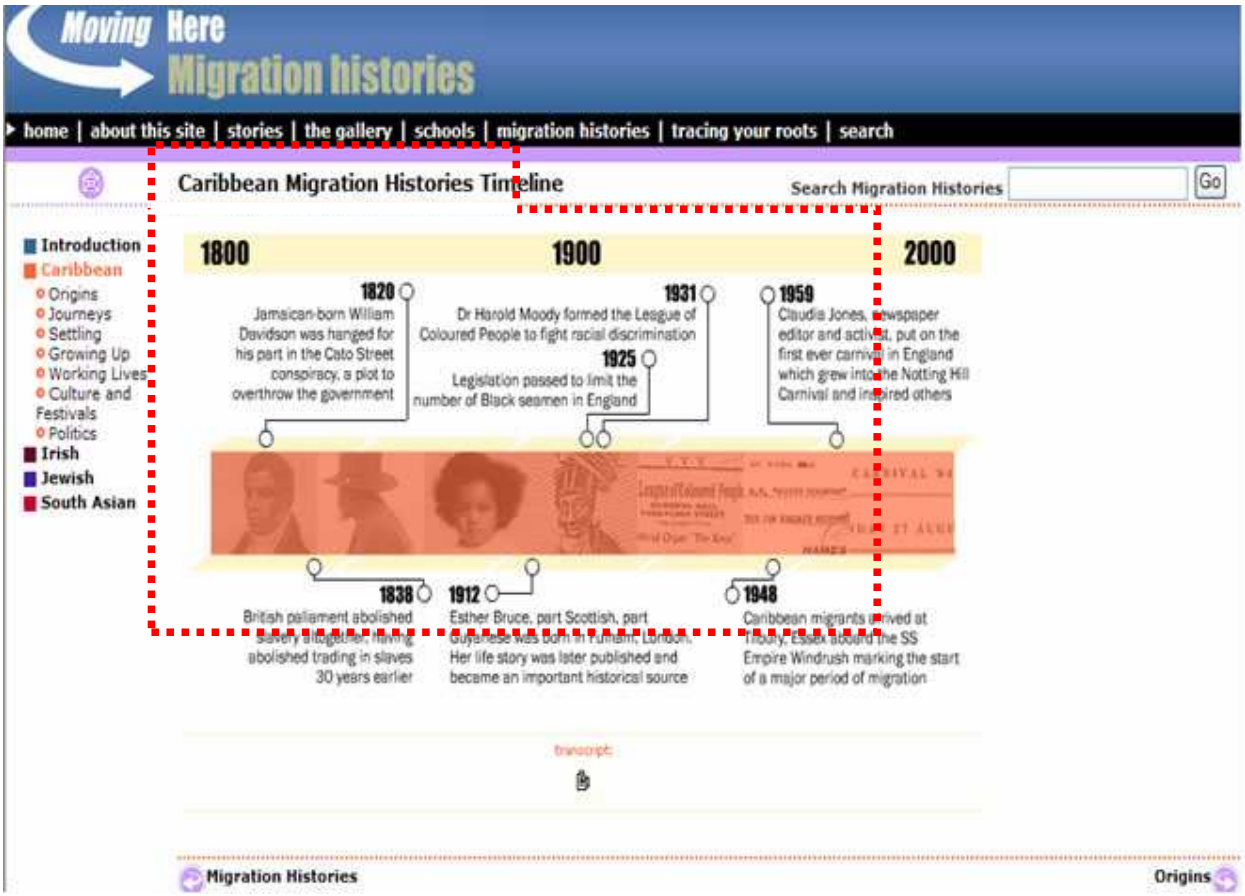

Fig. 17 – The use Pagelets in the Moving Here Website

8.1.3. Reconstructing hierarchical processes using a manual annotation tool

Website annotation systems designed to detect opinion and bias and changes in opinion and bias over time need to be based on hierarchical and multimodal principles. The previous section has established that semiosis (i.e. meaning-making processes) in Websites prioritise these principles over more traditional principles of literacy such as linearity and sequentially organised language structures such as paragraphs. A manual annotation tool, with some semi-automatic features, the *MCA Web Browser*, is currently being developed to annotate Websites relating to the various thematics explored within the project. It is designed to speed up the process of multimodal genre analysis which in the current stage of research is a laborious process. As Fig. 18 shows the tool takes the form of a Website browser capable of *annotating the different hierarchical levels found in Websites* in terms of a series of coloured rectangles according to position in the hierarchy: the current scale posits the existence of 5 levels of structure in Web pages which, from lowest to highest, are: *SubClusters, Clusters, SuperClusters, Pagelets, Page*.

In the current stage of research the annotation process works from the lowest level upwards. This is because we know much about lower textual levels but very little about higher levels which in the current stage of research remain to be identified. Thus, in this system, language is not considered a higher level resource but rather as a very low level resource used in combination with other resources to build higher structures such as the *timelines* we have analysed in the previous section.

Together with language, we need to consider visually oriented structures such as shapes and lines. Kress and van Leeuwen [53] hold that shapes are the visual counterparts of nouns in that they represent Participants, while lines represent Processes and are thus the visual counterparts to verbs. We posit that lines have major functions in Websites e.g. the pointing (or deictic) function of lines such as the *Swoosh Logo* in Fig. 13. We may note that framing is also a very important feature of Websites carried out by lines. An example is given in Fig. 16 where four *Photos* are used to frame the abstract concept of a visual repository. It is represented in the centre by an *Icon* which could not have this meaning if it were

not contextualised by the surrounding frame. *Frames* are thus important devices in the co-contextualisation of resources and a further evidence of the hierarchical nature of meaning-making processes in Websites.

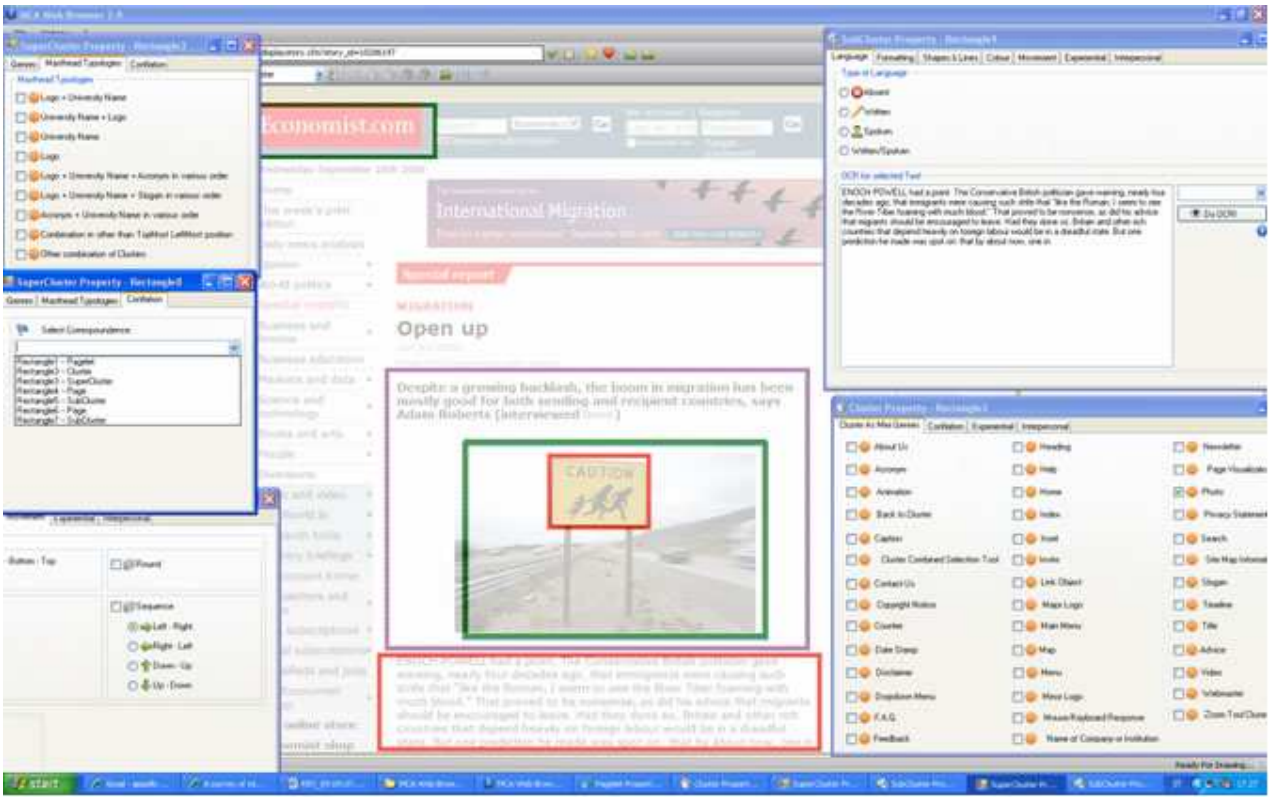

Fig. 18 – MCA Web browser

Although some Websites contain no dynamic objects – from a semiotic standpoint – movement is a typical characteristic of today's Web pages and is of crucial significance in many Websites, for example, those relating to climate change (e.g. in *animations* that simulate future changes in climate). Notice, however, that the number of parameters to be entered decreases as we move up the hierarchical scale. For instance, *Clusters* function as *mini-genres.* That is, they are the basic objects vis-à-vis which we are literate and which we immediately recognise as we move from one Website to another. We can posit that *SuperClusters* are at the level of the ***thematic system*** where more abstract meanings are made. Since it is the level where opinion-forming is carried out, we need to know more about it. A simple example of a thematic system relevant to the current research is the climate change conflict between the "non interventionists" and "interventionists" (i.e. bias in the assumption that climate change is manmade or not). Typically, such a conflict is expressed, for example in an online journal, in terms of a specific instance such as a photo of a coal-fired power station belching out smoke and a verbal caption of the type "Is this the legacy you wish to leave for your children?". The combination of photo + caption is a relation between two clusters which form a *SuperCluster*. It is the linkage between the two clusters (technically a covariate tie see [48]: 139) that makes it possible to index (i.e. "evoke", "have in mind", "think of", "be aware of") the underlying thematic system. ***Thematic systems are typically intertextual and more abstract than the specific meanings of specific texts.*** That is, they are implied meanings. The overt meaning in this hypothetical example is something like: this is a bad example of pollution; we shouldn't have power stations like this; the implied meaning is something like: a conflict exists between interventionists against non-interventionists and you, the reader, are being recruited as a potential interventionist. This latter meaning is intertextual, that is we build it up from our experience of many texts on similar thematics; we do not build it up from a single text.

On the Web, such thematic systems are typically intersemiotic (i.e. multimodal) as compared with many but by no means all printed texts, whence the role of multimodality and, in particular, multimodal annotation of Web pages in the current project to discover how such systems with their underlying

ideological potential (bias/persuasion) can be detected. Typically, such thematic systems will be linked to similar systems by partially overlapping with them (e.g. the potentially associated conflicts between young/old generations, new/old technologies, left-wing/right-wing political affiliations). *This stage of the research will provide some basic input for the understanding of how intertextual thematic systems are realised in terms of textual objects on a Web page in a corpus of relevant Websites.*

Thus, it is important to keep the centrality of intersemiotic thematic systems in mind while building the prototype. Some of them will be strongly ideological in nature while others will be ideologically weaker. An example of weaker ideology is the relationship between top and bottom bars on Web pages. The *Top Bar* is a *SuperCluster,* typically made up of a series of mini-genres which typically proposes the *Ideal*: a perfect service/product (think of a University Web page, or a car Website); the *Bottom Bar,* instead, in most Websites contains *Real* information i.e. information in the form of clusters and mini-genres which do not occur elsewhere e.g. disclaimers, legal notices, copyright, Webmaster notices etc.

It will be important to mark cases where expected SuperClusters are present or missing. For example, it will be important to consider *Real/Ideal* as a thematic system and to understand what effects there are if one part of the equation is missing. "Empty *SuperCluster* annotation" is the case where a *SuperCluster* is explicitly marked in annotation/detection systems such as the *MCA Web Browser* as absent and which help in the process of explaining why a particular configuration or pattern of objects is present or absent in certain circumstances.

8.1.4. Recovering typical meaning-making patterns and identifying multimodal grammar

As stated in the first paragraph of this section, meaning-making patterns in Websites are recoverable through database search facilities. These characterise the existing UniPV corpus annotation system which is currently being extended and integrated with the *MCA Web Browser.* It is expected that the system will be able to provide answers to at least some of the questions listed in Appendix B. These question-probes constitute a framework for the description of a multimodal grammar based on semiotic principles.

8.2. Macro-strategies to detect opinions and bias in Websites

This section lays the foundations for research and development vis-à-vis a multimodal semiotics approach to Website analysis. Whereas many previous sections deal with micro-strategies based on *language* for the detection of diversity in opinion, this section deals with macro-strategies and assumes that meaning-making processes in Web pages work within a *hierarchy* of semiotic (i.e. meaning-making) units. The section also develops a strategy for the identification of diversity in opinion-making in terms of (inter)semiotic *subpage, whole page* and *cross-page* annotation and retrieval methods. It further assumes that, within this hierarchy, rather than at lower levels, meaning-making processes, and in particular interpersonal meanings, such as opinion and bias, are typically (though not exclusively) enacted at higher and more abstract levels thanks to genre patterning (in particular at the *Page, Pagelet* and *SuperCluster* level). One consequence of this is that, rather than to focus on images and written text in *isolation,* it is appropriate to consider patterns of mutual ties and *interactions* between them which are often made 'explicit', for example, by devices such as *encapsulating frames* and/or *captions*. The final suggestion is thus that these higher and more abstract levels are potentially identifiable and retrievable with automatic detection methods as work with the *MCA Web Browser's* whole page annotation and *MCA*-corpus based methods [45][46][43] bears out.

8.2.1. Theoretical stance of multimodal semiotics

A multimodal semiotics approach views texts (where the term "text" means a semiotic unit such as a Website but also an object such as cup of coffee or a fire extinguisher) as *dually* material and semiotic entities. Processes, in which the *material* and the *semiotic* dimensions are the two sides of the same textual coin, are fully integrated into the same overall contextualising activity ([48]: 175). Even though action *is* important in a semiotic approach to Websites, the focus here is necessarily on the *hierarchical organisation* of a Website as "text" (i.e. as a patterned multimodal composition/meaning-making unit)

and, in particular, as the material expression of a hierarchy of genres that instantiate expected compositional and actional patterns.

Websites are multisemiotic, hierarchical structures: Websites are dually meaning-making textual/compositional and actional frameworks that enact basic meaning-making processes, such as informing, persuading, warning, criticising, appraising, and so on. Language-based studies alone cannot exhaustively represent the semiotic structure of Websites. The reason is that Websites respond, *inter alia*, to *topological* (i.e. spatial) principles of organisation as well as *typological* (i.e. categorising, language-based) principles. Analysis of today's Websites indicates that they are very different from Websites of 10 years ago. They exhibit both: action-based structures and *prefabricated multimodal compositional structures* in which *the incidence of fixed elements is likely to increase* with progressing time. Today's Websites are typically:

- **Intersemiotic (i.e. multimodal) rather than language-only structures**: they rely on integrated visual, spatial and linguistic resources rather than just on language in the creation of meaning; like many *home pages*, the Web page in Fig. 13 contains no paragraphs and indeed no sentences of the type typically associated with paragraph-based *running text* (cf. *newspaper articles*) characterised by explicit subject, verb nuclei and predicates. The only sentences used are subjectless (and hence themeless) imperative forms: "Find out...", "Go to...", "Contact us…" and so on; instead meanings are mostly made through highly elliptical linguistic structures which are intertwined with visual and spatial resources
- **Hierarchical and cyclic rather than linear**: many Websites (e.g. Fig. 14) encase written texts in explicit frames that guide the page-scanning and reading process. Websites base their thematic expansion on periodicity and visual/spatial subordination. Frames are indicative of a hierarchical scale of page subunits running from *page* to *resource/subcluster level* via the following sequence *Page>Pagelet>SuperCluster>Cluster>Resource/SubCluster level*. In this regard, language is assumed to be a low-level *resource* which only instantiates meanings at higher levels through its combination with other resources. Many of these combinations, most obviously language, orthography and colour, have undergone remarkable evolution in Internet's short history
- **Meaning-multiplying and meaning-compressing**: the integration has a multiplying rather an add-up effect. Today's Websites accordingly make more meanings per page than those of 10 years ago. This increased capacity to mean (meaning potential/meaning density) relies on the increasing use of prefabricated structures (*multimodal collocations*) which are constantly recycled and customised
- **Pattern-forming**: from one standpoint, the Web is a hierarchy of nested genres. After the first introduction of a new genre an experimental phase typically takes place in which the new genre either dies out or catches on.. In the latter case, it gets copied and a process of standardization takes place, which increases the *predictability* of the characteristics, and to some extent, the incidence of the genre. For example, 10 years ago *mastheads*, the ways in which companies, institutions, associations and other bodies name and identify themselves, were far less predictable than they are today. For University *mastheads* we can draw up a 'rule' which states that, today, universities identify and name themselves through a *masthead* complex which consists of a *logo* symbolising the university and a *name* providing the name of the University in the form of *University of X* or *X University*. This complex *will be* located in the top *left hand corner* of a *home page*. The University's logo *will be* present and *will have* the form of a *coat of arms* or *shield* (as opposed to a cup of coffee or a fire extinguisher). It *will be* positioned to the left rather than to the right of the University's name. Based on a small corpus of 30 University Websites in English language Websites, of a set of 16 possible masthead types taken into consideration, this 'rule' works for over 40% of examples analysed for 2009 but less than 14% for a corresponding set of examples from 1999 when the presence, composition and positioning of the *masthead* complex was much less predictable; of 16 possible masthead types posited, 8 existed in 2009, while 11 were attested in 2004 and the number rises proportionately as we move back in time. We may think of the Web as a set of prefabricated moulds in which the less successful get rapidly discarded. As the Web grows some of its genres and underlying configurations are

becoming easier to predict as they follow norms that result in pattern formation; in other words we can accurately track some aspects of semiotic evolution of the Internet through annotation and corpus construction, where "accurately" includes the provision of statistical data;

- **Subject to pressure-to-conform norm**: Website objects are becoming more and more standardised. This process is already increasing the predictability of Web page objects. The growth of more sophisticated search engines will further this process: were searches with *Yahoo*, *Google* etc. to visualise the *Top Bars* of Web pages, rather than just *Addresses*, the reaction by companies and institutions would be to further the already strong image-identifying functions of *Top Bars*, entailing a further rise in the predictability of the type of information and type of attitude found in Websites and their parts, thanks to the formation and crystallisation of new genres
- **Holistic, dynamic, trajectory-based and negotiational**: in terms of their meaning-making potential, Web pages are not like printed pages insofar as they are constantly reshaped by the dynamic elements in the page and by user interactions in the form of writing and selecting. When interacting with the whole page, users must select and negotiate which parts and functions to prioritise in order to build trajectories that will lead them to other pages. Accordingly, interaction with Web pages is no longer just a question of *hypertext links and selections* (i.e. reading and then clicking on a blue link for more in-depth information on a specific point). Instead, ***Website interaction increasingly resembles conversational interaction***. For example, it presupposes dialogic interaction based on *turn-taking norms* (cf. BLOGS) in complex decision-making processes which have notable consequences vis-à-vis the increased potential to express interpersonal/attitudinal meaning (i.e. opinion and bias)
- **Integrate interpersonal, ideational and textual meanings in predictable and hence potentially retrievable ways**: Halliday's social semiotic theory [54] of the intertwining of three types of meaning in all texts (ideational = e.g. experience of facts, events, thoughts, logical structure in their expression; interpersonal: e.g. attitudes, opinions; textual: semiotic form) and how these three types of meaning can be analytically separated in language-based discourse is a culminating point in many centuries of thinking about how meaning is made. The same model has been successfully extended to multimodal discourse with the result that statements to the effect the Web is 'anarchic' can be offset by the observation that typical meaning-making patterns in Websites are recoverable through manually annotated corpora and potentially through automatic means (Appendix C)

Multimodal genre analysis provides a high-level perspective on meaning making in Websites. Like all genres (i.e. clearly identifiable compositional patterns), multimodal genres establish a set of expectations about content and facts as well as attitudes to, and perspectives on, content and facts. They provide important clues to, and constraints about, what will be expressed in a Web page. Web pages, taken as macro-structures (rather than the various parts of their overall structure), are the *material means* that create expectations about the types of facts and opinions that will be expressed in Web pages and Websites in terms of the forms they will probably take. As such, they set up likely reading priorities and govern the way a user will extract information from the Web page.

8.2.2. Corpus-based approach to generic features in Websites

By providing quantitative and qualitative data, corpus techniques [49] based on multimodal and experimentation with eye-tracking techniques provide a halfway house in the leap from general principles to automatic retrieval techniques since they reveal basic patterns about genre sets and user selections from these genre sets. A multimodal corpus of Websites [45] annotated at the Webpage macro-level is thus a way forward in the project. Macro-level annotation, however, needs to be multimodal. While linguistic annotation methods based on mediated discourse principles typically apply to micro-levels (e.g. the language and discourse structure of reports). By definition they do not address the higher levels of Website discourse organisation characterised by the emergence of new meaning-making units unique to the Internet and which only in part derive from previous traditions of literacy and

which were expressible partially only by non-digital technologies. For this reason, care is being taken to ensure that the various methodological approaches are complementary. This matter also extends to the construction of (multimodal) LK-compliant corpus data that are 'readable' by the various tools being developed within the project. Exemplification of multimodal macro-annotation of a scalar type (i.e. a hierarchy of interacting genres ranging from *genrelets* to *macro-genres*) is given in Appendix C in relation to one Web page genre (namely on-line journals). It can also be applied to other Website genres such as Website newspapers, company Websites, government and institutional Websites, albeit with different results.

8.3. Conclusion

It is not by chance that today's Web is called *Web 2.0*. The new denomination reflects intertwined technological and semiotic changes that symbiotically fuel each other in a way that has qualitatively and quantitatively accelerated processes already existing in *Web 1.0* ([52]). The implication is that attempts to find a solution to the problem of developing automatic detection systems designed to identify diversity, opinion and bias, and changes thereof over time, can benefit from "atomic" *whole-page* procedures to full Website annotations. Web pages are semiotic structures whose meaning making takes place at *page level* as well as at *cross-page level* and at *sub-page level*. As suggested *inter alia* by eye-tracking studies (see Fig. 23 in Appendix C), it is the hierarchy of interacting semiotic structures that produces meaning for users in Web pages. This example provides experimental support for the principle that images and written text are not 'read' in isolation, but are part of a higher order meaning-making structure which controls the interactions with, and between, its subparts.

As the Internet evolves, it is increasingly dominated by compositional hierarchies. There is a decrease in the number of *pagey* Websites that rely on the parchment-based principle of scrolling and a corresponding increase in the 3-D properties of *screeny* Web sites (e.g. Fig. 13) which rely on horizontal organisation and 'piercing-the-page' access to information. Central to the semiotics view of Websites described here is the co-contextualising nature of Website objects that derives from the semiotically hierarchical organisation of Websites. It is this co-contextualising property that determines the function/meaning of each Website object within a Web page and which governs the ways in which users extract, and act on, Website information.

9. Contributions to the solution from library and information science

This section describes the basic notion of Faceted Classification (FC), construction of a FC and document classification using the techniques of Facet Analysis (FA). A corresponding glossary can be found in Appendix E.

9.1. Introduction to Faceted Classification

A library classification is a system of coding and organising library materials according to their subject. A classification consists of tables of subject headings and classification schedules used to assign a class number (or notation) to each item being classified based on that item's subject. Traditionally, three main types of classification are thought to exist [70]:

- **Enumerative**
- **Hierarchical**
- **Faceted or Analytico-Synthetic**

The purpose of each type of classification is the same [77]:

1. to normalise the language of documents on one hand and the language of questions on the other
2. to serve as a useful device to the indexer in the intellectual task of characterising the subject contents of a document (to display synonyms, hierarchical and other relationships, to facilitate intelligent selection of terms by the indexer)
3. to provide a tool to the searcher in analysing and defining questions to be put into the file

The method by which these purposes are accomplished differs for each system. With enumerative classifications each subject is subdivided until all possibilities are exhausted. Each subdivision is assigned a notation, or call number. In a hierarchical classification, each class is subdivided and subsequently ordered, placing the most general first and then moving to the most specific. Analytico-Synthetic (or faceted) classification developed in order to provide an alternative to what was often called the "Procrustean bed" of the enumerative and hierarchical classifications like the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC). A faceted classification divides subjects into mutually exclusive, orthogonal categories using the technique of facet analysis [79]

9.2. Facet Analysis and Faceted Classification

A faceted classification is a schedule of standard terms[23] to be used in the subject description of documents. The terms are first of all grouped into homogenous subject fields (a.k.a. domains). Within each subject field the terms are divided into groups known as "facets". Within each facet, terms may be arranged hierarchically. This process of creating the FC schedule is called Facet Analysis (FA).

Clarity of ***'Facet' definition*** is important and consensus of definition among various practitioners is desirable, though difficult to reach in practice. The term 'Facet' is most often used as synonymous to: *category, attribute, class, group, concept, and dimension*. Ranganathan initially used the phrase "*train of characteristics"* while emphasising that facets, "inhere in the subject [entity] themselves, whether we sense them or not.". In describing facet, he stated that "*a classification of a particular universe [of entities] is made on the basis of characteristics"*. In this sense, a characteristic is a parameter (representing the principle used for the subdivision). Each parameter creates a dimension which usually falls into a small number of groups. Each group becomes a facet, and is itself multidimensional [70]. Furthermore, a facet is defined as "*a generic term used to denote any component of a compound subject, also its ranked forms, terms and numbers*" [70]. Groups of terms derived by taking each term and

[23] In library classification, standard terms are classification scheme specific and each scheme has its own controlled vocabulary. So these standard terms are classification scheme specific as there are no unanimously agreed institutional bodies for maintaining standards for library classification.

defining it, *per genus et differentiam*, with respect to its parent class [76]. In simpler words, it can be defined as "a homogeneous group or category derived from the universe of entities or knowledge by applying a set of characteristics". It may also be seen as one part of a subject, situation, etc. that has many parts.

Facet Analysis (FA) and Faceted Classification (FC) are not synonymous. Ranganathan [72][73] describes Facet Analysis as, "the mental process by which the possible trains of characteristics which can form the basis of classification of a subject are enumerated (*the process of analysis)* and the exact measures in which the attributes concerned are incident in the subject are determined". Analysis of a subject into its facets according to the postulates and principles stated for that purpose [70]. The essence of facet analysis is the sorting of terms in a given field of knowledge into homogeneous, mutually exclusive facets, each derived from the parent universe by a single characteristic of division [76].

Vanda Broughton [62] advises that, "Although faceted classification is regarded by many as a structure with specific characteristics, essentially facet analysis is a technique, and different models of the same universe of discourse can be derived to meet different local or subject-specific needs using different categories and variations on the syntax". Generally, according to J. Mills the facet analytical approach follows these steps:

Facet Analysis (FA) > Faceted classification (FC) > Facet Analysis (FA)

1. Identification of a universe of entities to organise
2. FA of a representative sample of entities and division of the entities into arrays
3. Once FA is complete, it is possible to create an FC which involves the formulation of filing and citation order, index and notation; it is not possible to create an FC without conducting FA
4. The process of the subsequent FA then assists in the classification of entities by facilitating assignment of subject access points (subject analysis) and notation

9.3. Basic steps in the construction of a Faceted Classification

FA is "the basic operation in constructing a faceted classification" [76]. FA is a form of conceptual analysis that begins by studying the terminology and identifying the terms used. These terms form the raw materials for analysis. Each term is examined and a series of questions asked: What concept does it represent? In what conceptual category should this concept be included? What are the class relations between this concept and other concepts included in the same category? The resulting faceted schedule is a conceptual scheme, a structure in which terminologically expressed concepts have been organised [77]. Vickery [76][77] and Gopinath [65] outlined some basic steps in the creation of an FC and this process shows the interplay of FA and FC. These steps can be summarised as below:

(1) **Define the subject field:** This is accomplished by first asking, "what entities are of interest to the user group, what aspects of those entities are of interest." [77]. This step involves the investigation of the domain in hand as well as the purpose of the classification [65]. **(FA)**

(2) **Formulate facets:** Vickery recommends examination of a representative range of material that directly expresses the interests of the user group: reports, papers, comprehensive texts, glossaries, subject heading lists etc. This provides a list of candidate terms to use. While deriving the terms from the domain "a set of characteristics" can be applied to deduce the terms. **(FA)**

- **a.** Sort these terms into homogeneous groups known as facets.
- **b.** Each group is derived by "taking each term and defining it with respect to the terms that are the centre of interest in the classification." [77].

(3) **Structure each facet:** Each facet is amplified and structured. It is helpful at this stage to construct a hierarchical order for the terms collected within each facet. Even if no well developed hierarchy emerges, the procedure helps to coalesce synonyms, eliminate terms that

are collated with the wrong facet, and indicate gaps in the system. **(FA and preliminary construction FC)**

(4) **Create scope notes.** These notes will define terms that are unclear and provide instructions to users and indexers as to the meaning and use of each facet. **(FC)**

(5) **Arrangements of facets:** Decide how the facets are to be arranged among themselves. This will depend on the use: For post-coordinate use (as in a thesaurus), arrange into categories. For a pre-coordinate system like a catalogue, more thought must be given to the sequence of facets in the schedule and placing them in citation order. The order chosen should be that which is thought to be of most use to the person using the system. **(FA/FC)**

(6) **Fundamental Categories:** Fundamental categories can be seen as broader labels under which the similar facets are grouped. There are five and only five fundamental categories according to Ranganathan: Time, Space, Energy, Matter, and Personality [70]. But this notion of Ranganathan was not widely accepted and different researchers came up with diverse sets of fundamental categories. The CRG (Classification Research Group) found a list of 13 helpful: substance (product), organ, constituent, structure, shape, property, object of action (patient, raw material), action, operation, process, agent, space, and time. Barbara Kyle wrote of: natural phenomena, artifacts, activities, and 'purposes, aims, ideas and abstracts'. De Grolier suggested the 'constant categories' of time, space, and action, and the 'variables', substance, organ, analytic, synthetic, property, form, and organization. Fundamental categories are useful as a provisional guide in approaching the analysis of a new field […] provid[ing] an outline framework […] and giv[ing] guidance in suggesting possible characteristics which should not be overlooked [76]. This topic needs further discussion and research so as to decide the suitable list of fundamental categories for almost all the disciplines of the universe of knowledge.

(7) **Create notation:** In this step notation is provided for each of the terms so that call numbers can be constructed for materials in a system classified using FC. This topic has been omitted from this primer, as the present work is not going to use notation. **(FA/FC)**

(8) **Fitting a notation:** This is the final result of FA, in addition to a schedule of terms and one of the ways in which the full sequence of structured facets may be displayed. This arrangement should display the structure of the subject field helpfully. **(FA/FC)**

In each of the above-mentioned steps, postulates, principles and canons of classification proposed by Ranganathan [70] are used as a set of guiding principles. The forthcoming report will enumerate all those relevant principles in building a practical classification scheme.

9.4. Classification of documents according to Faceted Classification

As is apparent from the previous sections, the FA plays an important role in preparing the classification schedule as well as in classifying the actual document according to the given schedule. This section elaborately deals with the processes and steps involved in document classification. The underlying basic approach involves breaking the given title[24] into various parts and then applying faceted analysis to place them under different facets. It is equivalent to translating the subject of the document from a natural language to the artificial language of ordinal numbers forming a classificatory language. The nine steps followed in succession while classifying a document are as follows (the steps are explained by using a hypothetical document with the following title and Ranganathan's five fundamental categories, i.e. Personality[P], Matter[M], Energy[E], Time [T] and Space [S] are used[25]):

[24] Here "title" refers to any string of words signifying a particular subject in the document.

[25] We suggest the reader to refer to [70] for a complete account of the fundamental categories.

Title: Nationality as an issue in forthcoming European elections

Step 0: ***Raw Title***
In this step the title found on the title page of the document is taken.

Title as found in the title of the document: ***Nationality as an issue in forthcoming European elections***

Step 1: ***Expressive Title:*** An expressive title is one which covers all the facets of the subject of the document. Often the Raw Title is incomplete and not fully expressive. Therefore, the following preparation is required to complete the Expressive Title:

i If the Basic subject (the domain) is not mentioned in the Raw Title, it is to be added in the Raw Title, as every compound subject has a Basic subject
ii If some isolate terms are absent from the Raw Title, then infer absent terms from the title-context and insert them
iii If the period covered by the document is not indicated in the title then the classifier has to find out the period covered by the document by perusal of the whole document and has to insert it in Raw Title
iv If the Raw title contains derived composite terms then these are to be replaced by fundamental constituent terms in order to make the Raw Title fully expressive
v If the Raw title is a 'fanciful title', then an expressive title is to be coined by the classifier after perusal of the document

On the basis of these steps, the expressive title would be:
In political science, nationality as an issue in forthcoming European elections to be held in 2009

Step 2: ***Kernel Title:*** A kernel title is one which contains only Kernel Terms. For this, the following preparation is necessary:

i Remove the auxiliary or apparatus words such as – in, of, by and up to
ii Indicate the kernel term in its nominative singular form
iii Change the first letter of each kernel term into a capital letter
iv Insert a full stop after each kernel term

The above title in Kernel terms would be:
Political science . Nationality . European . Elections . 2009

Step 3: ***Analysed Title:***

i Place the symbol [BF] against the kernel term denoting the basic subject
ii Find out the fundamental category for which the idea denoted by each kernel may be deemed to be a manifestation and place it against each symbol in each fundamental category
iii Determine the Rounds and Levels to which each Kernel term may be assigned
iv Assign the symbols representing Rounds and Levels to the kernel terms

The above title in Analysed Title terms would be:
Political Science (BF). Nationality (P). European (S).). Elections (E). 2009 (T)

Step 4: ***Transformed Title:*** Arrange the kernel terms along with their respective symbols in the correct sequence. The sequence is determined by using the principle of facet sequence [BF][P][M] [E][S][T].

Political Science (BF). Nationality (P). Elections (E). European (S). 2009 (T)

Step 5: ***Title in Standard Terms:*** Replace each kernel term in the title by its equivalent terms as given in the schedules of the scheme for classification in use. For example the above-mentioned title will be transformed into the following:

Political Science [BF]. Nationality [P]. Election Method [E]. Europe[S]. 2009 [T]

Step 6: ***Title in Focal Numbers:*** Replace each standard term by its class number or isolate number as is given in the schedules of the scheme. For example the above titles in focal numbers using CC6th edition would be:

W[BF]. 94[P]. 91[E]. 5[S]. P09[T]

Step 7: ***Class Number of the subject:*** This is the final step in which the name of the subject is translated into its class number in the preferred classificatory language.

So, the final call number/class number of the document will be (here the notations used are borrowed from Colon Classification, 6th edition): ***W94:91.5'P09***

Step 8: ***Verification:***

i Facet Analysis of the class number: this is done by writing each facet number in the subsequent line along with their respective symbols.

ii Transformation into kernel terms: this is done by writing facet number in succession and against it in the Basic Subject or Isolate Term, as the case may be, which it represents.

iii Assembly in Skeleton Form: for this the Facet Terms obtained in the preceding section are to be assembled in the sequence in which the symbols for the facet indicate.

iv Transformation respecting the syntax of the English Language: rearrange the facet terms obtained in the preceding section in accordance with natural language syntax, inserting the necessary apparatus words.

v Verify whether the name of the subject reached under the preceding section is equivalent to the raw title of the document being classified.

vi Rectification: If after verification, it is found that the name of the subject reached is not equivalent to the document's raw title, it means that at one step or the other, some mistake has been made. Trace it out and rectify it and verify again.

9.5 Conclusion

A representation based on faceted ontologies (i.e. ontologies built from FCs, e.g. [215]) of information in a generic form would lead to semantic retrieval. The knowledge structure to be deployed for the purpose corresponds to a semantic framework that is really a generalisation or abstraction of formal representation of domains [80]. Each domain is envisaged as consisting of divisions or pieces of knowledge called facets where a facet is a distinctive division of the domain or subject that is conceptualised. Each facet in turn contains a set of concepts of the domain in a hierarchy and many such

facets together comprise a subject domain. This model is based on Ranganathan's theory of facetisation [70]. The generic manifestations of such subject representations lead to faceted ontologies.

A formalised hierarchical structure of concepts forms a facet. All such facets are assembled together as required to describe information in each scenario. The facets are recognised as entities, action and properties of the entities [78].

While the facets themselves are distinct divisions of the domain and contain the concepts belonging to the facet within them, there are rules to generate surface strings for any given term that bring along with them the context. These are formalised strings that represent concepts in their entire context generated for tracing a term or its variant form occurring in various domains, by using certain rules in the system about how each term is related to the other and in turn the relations between the resources [81].

10. Interplay between technologies and methodologies

This section mainly discusses current cross research activities between technical and methodology partners. All of them are involved in the process with their individual contributions. Many more cross partner activities are envisioned for the future. The future work will mainly consist of addressing the issue of representing and managing opinion, bias, diversity and evolution in more detail. In fact, it should be noted that much of the planned future work on basic text and image analysis technologies relates to the techniques covered in the state of the art section. In this section we address in particular the future work relating to the interplay between the technologies and the methodologies (described in Sections 7, 8 and 9). We will give particular importance to opinion and bias detection, the spatial and temporal dimensions of knowledge as well as the notion of facet. Of fundamental significance will be the investigation on how technologies can be used to automate the annotation processes described by the single proposed methodologies. We currently plan to proceed at least in the following four directions.

Extraction of the variables to fill the Codebook

The Media Content Analysis (MCA) approach of SORA (described in Section 7) involves the use of a Codebook listing variables and indicators, which are – up to now – extracted manually from documents. Technical partners have analysed the Codebook to investigate which variables can be extracted automatically, semi-automatically or manually. Even more, additional variables that might be useful for the annotation process have been listed by the technology partners. The results of these analyses are part of the deliverables in WP2 and WP8. As a preliminary step, the generation of a gold standard from a document corpus (generated in the context of WP8) is necessary. These documents will be manually annotated using the already mentioned methodologies from SORA, ISI and PAVIA.

The Codebook variables used in MCA are currently extracted manually by analysts. Clearly, an automatic tool that extracts Codebook variables would be valuable for media content analysts. However, to develop an automatic extraction tool, a collection of manually annotated data is needed to train statistical extractors and to evaluate their performance.

For this purpose, UNITN has been collaborating with SORA to annotate the Codebook variables as they appear in news text. The Callisto annotation tool [216] has been selected, and the design of annotation guidelines is underway. UNITN configures the Callisto tool for the annotation task at hand. The annotation of a large corpus of text will be beneficial not only for the purpose of creating tools for MCA, but also for the field of automatic opinion and sentiment analysis in general. An interesting methodological question is in what way this corpus will differ from previously published opinion-annotated corpora such as the Multi-Perspective Question Answering corpus (MPQA) [217], which focuses on annotating the surface-linguistic expressions of opinions.

Automatic systems that extract and label pieces of text are today almost exclusively implemented as statistically trained sequence labellers (at least when annotated training data is available). This is a very well-researched area, and an off-the-shelf sequence labeller such as YamCha [218] will serve us as a high-performance baseline. However, since the classical sequence labelling formulation uses very limited contextual information, we will explore more sophisticated methods based on more complex feature interdependencies. This obviously makes exact inference computationally intractable, and we will explore re-ranking as an approximate search strategy, which has been proven successful in several problems in natural language processing such as parsing [219] and semantic role labelling [220]. In re-ranking, a small candidate pool is generated by a simple model (in our case a conventional sequence labeller), and the complex model selects the top-scoring candidate from the pool rather than from the complete output space.

In addition, UNITN will explore the connection between opinion structure as annotated by media content analysts and the discourse structure of the texts. Discourse structure extraction algorithms will be trained on the Penn Discourse Treebank [221]. Automatic discourse structure analysis consists of a number of diverse sub-problems that will need to be tackled separately. For instance, we believe that for sentence-internal relations between explicit discourse connectives and their argument, semantic role

labelling techniques that rely on syntactic structure may be adapted. On the other hand, the classification of discourse relations between adjacent sentences cannot rely on syntax and is still to a large extent an unexplored area. We will try to adapt kernel-based text pair classification methods [222] to tackle this problem. In addition, we will study if the prediction of discourse structure can be improved by some sort of global model, such as a sequence model.

ISI's Facet Classification

UNITN and ISI are working on ways to integrate the ISI's Faceted Classification with Trento's work on lightweight ontologies. The result is what we call Faceted Lightweight Ontologies [215]:

> *A **faceted lightweight ontology** is a lightweight ontology where each term and corresponding concept occurring in its node labels must correspond to a term and corresponding concept in the background knowledge, modelled as a faceted classification scheme.*

See the example in Fig. 19 and taken from [215]. Notice that here the fundamental categories are those used by Bhattacharyya [78]. All the terms occurring in the labels of the faceted lightweight ontology on the right correspond to a term and corresponding concept in the medicine domain background knowledge. They have a well defined structure and, as such, they are easier to create, to share among users, and they also provide more organised input to semantics based applications, such as semantic search and navigation. In fact, structures commonly used by users to classify their material (documents, photos, music …), such as Web directories, folder hierarchies in a file system, email folder structures and so on, can be easily (and almost without loss of information) translated into faceted lightweight ontologies.

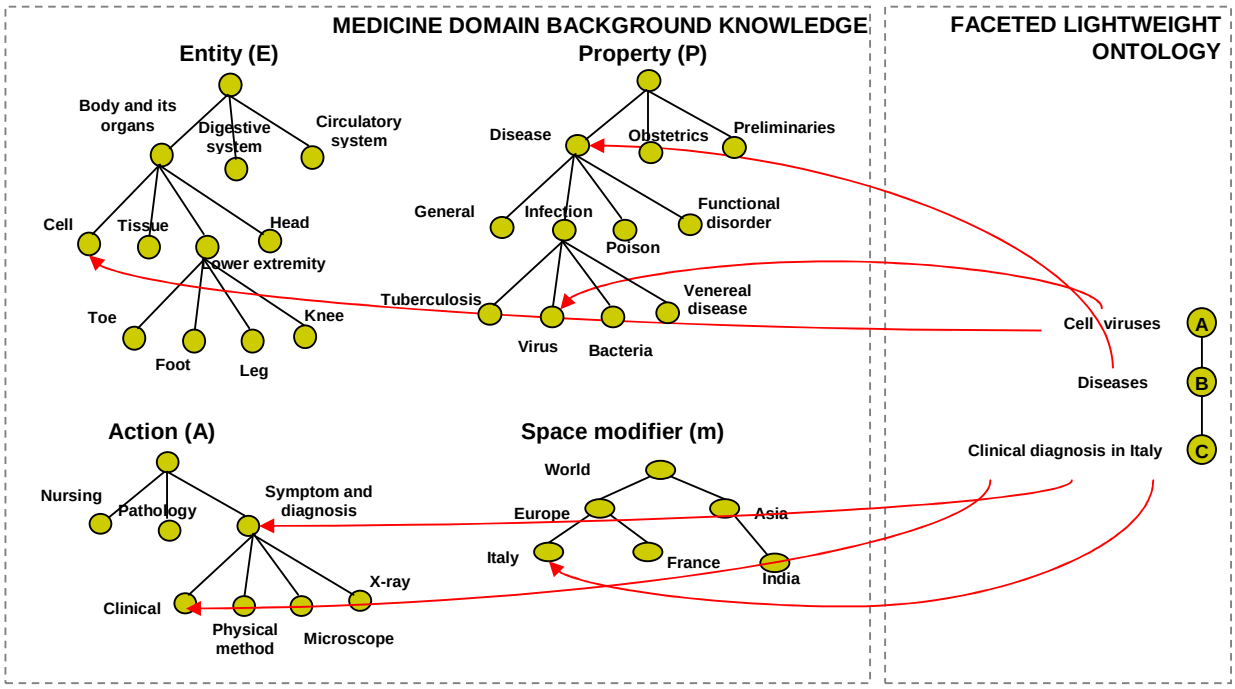

Fig. 19 – An example of faceted lightweight ontology grounded in the medicine domain background knowledge

Text processing and the SORA/ISI/PAVIA methodologies

As mentioned in the state of the art section, the main techniques being investigated for text mining involve mainly extraction of facts and opinions, identification of dimensions and sources of diversity, classification of documents by topic and genre, as well as summarization and aggregation of documents.

According to the MCA methodology (by SORA), the starting point of the process is the specification of a *research question*, for which information has to be collected, analysed, and presented. This is made

more concise by a process of extracting values for specific *variables* related to the research question. Hence, this step corresponds to the query formulation and disambiguation process. Starting from an information need, a query needs to be formulated, which can take various forms, such as a set of keywords, a structured query, or a natural language query. Furthermore, the query terms may be ambiguous, which leads to an additional dimension of diversity[26].

In the next step, relevant information from various sources has to be identified and extracted. This should cover a variety of factors that possibly contribute to the diversity of the information related to the research question. A basic issue is to identify and analyse the different topics, as well as their possible sub-topics, that may occur in the selected set of documents. This problem can be addressed using different classification and clustering techniques. Several tools (mentioned in the state of the art section) already exist to facilitate this task. Another important aspect is the classification of sources with respect to genre, register, and type. This analysis can reveal the cultural and situational context used in the documents, which is important for analysing diversity and bias. Multimodal Genre Analysis (MGA, by PAVIA) may be helpful for indicating guidelines for this task, as well as pointing out on which parts of the documents to focus for the extraction process. Another critical task is the extraction of facts and opinions. Once relevant statements have been extracted from documents, methods have to be applied to distinguish between objective and subjective statements. Extracted opinions need to be further classified according to their polarity. This is usually addressed as a three-class classification problem, namely classifying opinions as positive, negative, or neutral. However, techniques for more fine grained characterization are being investigated. This task can reveal what opinions are expressed about the identified topics or subtopics, and furthermore identify similar or contradicting opinions, as well as how these opinions evolve over time. Finally, additional metadata can be extracted, as explained in the analysis of SORA's codebook[27], involving the opinion holders, or information about time and location of identified events. MCA can provide guidelines regarding the information to be extracted from the selected set of documents. Due to the need for a sufficient amount of training data in order to apply machine learning techniques, the manual annotations provided by SORA, ISI, and PAVIA will be especially important.

Once the relevant information has been extracted and analysed, the creation of faceted classifications is needed to facilitate better exploration and navigation of the results. The main challenge that arises here is how to automatically identify interesting facets. Usually, this task is performed manually, which requires that the categories of interest are known in advance. In addition, this is a time consuming process, and important aspects in the data may potentially be overlooked. Hence, the need arises to automatically (or semi-automatically) identify interesting facets. For example, existing work attempts to identify facets that are good descriptors, i.e., they are most effective for representing the dataset, as well as facets that are good navigators, i.e., they are most effective for browsing the dataset. Facet Analysis (FA, by ISI) can provide useful guidelines for defining metrics for facet selection. This work is related to ISI's and UniTN's collaboration on faceted classification and lightweight ontologies. Finally, the results may need to be summarised and aggregated, which is typically achieved through clustering techniques.

Image processing and the SORA/ISI/PAVIA methodologies

As previously mentioned, many features can be extracted from images in particular and multimedia in general. An interesting aspect of the image research in the project is to explore if, in particular circumstances, particular image features are associated with particular opinion sets, biases or aspects of diversity. UNITN and others in the project are exploring the research area of opinion mining and sentiment analysis, evaluating what features can be associated to different opinions and emotions conveyed by images. In cases where there is an association, multimedia analysis may help to clarify ambiguities from text analysis. Large test sets of documents classified into opinion classes or bias/diversity groupings and containing visual as well as textual information are required to progress this type of approach. Of course, the image classes may be generated automatically from documents

[26] These issues are extensively addressed in WP4 (see the report "Bias and Diversity: Approach, Methods, and Algorithms")

[27] See the reports delivered in WP2.

successfully classified using text alone and may then possibly be used to help to disambiguate documents where the bias is unclear from an analysis of the text only.

Images may often have associated metadata (e.g. exif data, captions, etc) and this may also prove useful for opinion mining in the same way as conventional text. SOTON's work on automatic image annotation (e.g. [211][212]) can assign keywords to un-annotated images but requires large training sets of annotated images in the same domain. By analysing predicted annotations from an image it may be possible to link it to segments of the surrounding text that contain similar terms, and to thus gain an estimate of the opinion or fact that the image is trying to portray or support from opinions/facts expressed in the text.

In addition, the use of large-scale robust image matching technologies [213][214] may allow the provenance of image data to be determined; whether or not the original source of an image corroborates with information about the image in the document being analysed may hint at bias in the document.

Automatic image content analysis might help with some of the stages in this process of using computer-based tools to accelerate manual annotations, as used for example in the multimodal analysis from UNIPV. For example, face detection may be used to discover the presence of faces although robust face detection is mainly only achievable on clear near full frontal faces. However, this may be useful and analysis of the face may possibly be able to distinguish between smiling and non-smiling faces in some cases. Recognising that someone holds a particular object (out of a wide range of objects) is much more difficult and would typically require specific object detectors trained on large numbers of images of instances of the objects.

The automatic linking of (automatically) annotated images with the surrounding text is something that it is possible to explore. SOTON aims to continue improving and developing automatic annotation techniques and investigate how linking of the annotated terms with the co-located text might be best achieved. Techniques to automatically annotate images are being carried out in UNITN as well. In particular, UNITN focuses on annotation techniques for extracting relevant objects in images. As introduced before, these annotated terms could be linked with the co-located text to obtain a better estimate of facts and opinions expressed in a document.

The intended source documents for the LivingKnowledge project are expected to be multimodal in nature. That is, Web-pages and documents that contain images and text, graphically designed such that they promote a particular viewpoint. Identifying regions of the page, which may contain pertinent information with regards to opinion and bias requires an approach like the multimodal analysis technique from PAVIA. To automate the multimodal analysis it is necessary to first identify potentially interesting regions of a document, or regions that are in some way linked (clusters).

One method to find these would be to use the hierarchical structure of the document’s code (HTML for Web pages, XML for certain documents) that represents the document. In many cases, related content in documents is grouped into paragraphs or structural elements like tables or page divisions. With some explicit rules we may be able to automatically detect these elements. In the worst case, it should be possible to provide a semi-manual mode where regions are automatically detected and the user selects which ones they feel are pertinent.

Another possible method to detect clusters is to create an image of the rendered document (effectively perform a screen-grab of the document in an automated way) and perform image analysis on it. Using this technique it may be possible to find clusters that span the underlying structural elements in documents where the structure is not easily detected. The accuracy of this method is likely to be lower as it would rely on image analysis techniques which are much less robust than using the explicit document code.

Once potential clusters have been detected, the content of those areas needs to be analysed. Back-projecting into the document model to find the content of the clusters and then performing image or text analysis on the contents may provide a way to determine if the regions are in fact useful and, if so, whether the clusters represent high-level or low-level content, whether the clusters are related to other clusters (e.g. containing similar logos, similar colours) and whether they might have specific inferences (e.g. smiling or frowning people pictures). Links between clusters may be determined through various contextual analysis techniques.

Some of SOTON's work is based not on extraction of image content, but an analysis of the context in which the content was created. For example, content analysis of an image of the leaning tower of Pisa might possibly be able to identify it, but if the image is geo-tagged, the possible content of the image may be inferred before even looking at it. The contextual information surrounding an article is also relevant. For example, an initial estimate of the bias of a newspaper article on a particular issue may sometimes be inferred by knowing its subject and source. An article in a right-wing publication is likely to have a completely different view on, for example, immigration than a left-wing publication. Analysing the provenance of the information in a statement is a key step towards determining not only the likelihood of it being an opinion as opposed to a fact, but also a way of inferring what the opinion might be, even before any content analysis needs to begin.

Contextual analysis can assist with media content analysis by simply providing a database of background knowledge on a range of relevant subjects. For example, if analysing a document on immigration the document may refer to other sources beyond what is immediately available. It should be possible to generate a digest of the document's sources in order to determine possible bias. If all sources are generally known to have the same views on the subject then the document probably does not present a balanced viewpoint.

In the context of facet classification, contextual analysis allows better classification based on existing knowledge. For example, an article on a historical figure is unlikely to specifically mention that the subject is dead, assuming the reader already knows this. A computer classifying articles relating to dead people may not pick this up without some kind of background knowledge base upon which to call.

References

References from socio-political Media Content Analysis

[1] Adam, Silke/Berkel, Barbara/Firmstone, July/Gray, Emily/Koopmans, Ruud/Pfetsch, Barbara/Statham, Paul (2002): Codebook for content coding of commentaries/editorials. Codebook from the Project: The Transformation of Political Mobilization and Communication in European Public Spheres. 5th Framework Program of the European Commission. Europub.com

[2] Bauböck, Rainer (1994) (ed.): From Aliens to Citizens. Redefining the Status of Immigrants in Europe, Avebury, Aldershot.

[3] Berry, John W (1997): Immigration, Acculturation, and Adaption. In: Applied Psychology: An international Review, 46, pp. 5-34

[4] Bosswick, Wolfgang, Friedrich Heckmann (2006): Integration of migrants: Contribution of local and regional authorities. Report for the European Foundation for the Improvement of Living and Working Conditions.

[5] Brantner, Cornelia/Dietrich, Astrid/ Saurwein, Florian (2006): Europäisierung der österreichischen Öffentlichkeit. Mediale Aufmerksamkeit für EU-Politik und der veröffentliche Diskurs über die EU-Erweiterung. Forschungsbericht, NODE Researchaustria/BMBWK. Wien: Institut für Publizistik- und Kommunikationswissenschaft. http://www.univie.ac.at/Publizistik/Europaprojekt/

[6] Donges, Patrick/Imhof, Kurt (2001): „Öffentlichkeit im Wandel". In: Jarren, Otfried/Bonfadelli, Heinz (Hg.): Einführung in die Publizistikwissenschaft. Berlin/Stuttgart/Wien: Haupt UTB, S. 101-133

[7] EAPC – European Association of Political Consultants (2005): Election Time. The European Yearbook of Political Campaigning 2004, Leibnitz.

[8] Esser, Hartmut (2000): Soziologie. Spezielle Grundlagen. Band 2: Die Konstruktion der Gesellschaft. Frankfurt am Main: Campus

[9] Esser, Hartmut (2006): Sprache und Integration. Frankfurt am Main/New York: Campus

[10] European Greens (2006): A Green Future for Europe. Adopted as amended by the Congress of the European Green Party Geneva 14th October 2006.

[11] Ferree, Myra Marx/Gamson, Willlian A./Gerhards, Jürgen/Rucht, Dieter (2002): Four models of the public sphere in modern democracies. In: Theory and Society 31/2002. pp. 289-324

[12] Gerhards, Jürgen/Offerhaus, Anke/Roose, Jochen (2007): The Public Attribution of Responsibility. Developing an Instrument for Content Analysis. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 59. pp. 105-125.

[13] Giugni, Marco/Passy, Florence/Statham, Paul (2005): Institutional and Discursive Opportunities for Extreme-Right Mobilization in Five Countries. In: Mobilization. The International Journal of Research in Social Movements, Protest, and Contentious Politics 10. pp. 145-162

[14] Giugni, Marco/Statham, Paul (2002): The Contentious Politics of Unemployment in Europe. Political Claim-making, Policy Deliberation and Exclusion from the Labor Market. European Political Communication Working Paper Series, Issue 2/02. Leeds: Center for European Political Communications.

[15] Gruber, Oliver/Herczeg, Petra/Wallner, Cornelia (2009): Integration im öffentlichen Diskurs: Gesellschaftliche Ausverhandlungsprozesse in der massenmedialen Öffentlichkeit. Analysiert anhand des Fallbeispiels „Arigona Zogaj" in den österreichischen Medien. Project report, University of Vienna

[16] Habermas, Jürgen (1992): Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts

[17] Haller, Max (2009): Die Europäische Integration als Elitenprozess. Das Ende eines Traums? Wiesbaden.

[18] Han, Petrus (2000): Soziologie der Migration. Stuttgart: Lucius & Lucius.

[19] Heckmann, Friedrich, Dominique Schnapper (2003) (eds): The Integration of Immigrants in European Societies: National Differences and Trends of Convergence. Stuttgart.

[20] Hofinger, Christoph, Harald Waldrauch (1997): An index to measure the legal obstacles to the integration of migrants, in: New Community 23(2) - Special issue: Incorporating Migrants in Multicultural Societies, Issues of Citizenship and Equity, 271-285.

[21] Hofinger, Christoph, Ruth Picker, Eva Zeglovits (2005): The Campaign for the European Elections in Austria: Skepticism Rules, in: EAPC: Election Time. The European Yearbook of Political Campaigning 2004, Leibnitz, 11-26.

[22] Klaus, Elisabeth (2006): Von der Beschränktheit unserer Öffentlichkeitstheorien im europäischen Kontext. In: Langenbucher, Wolfgang R./Latzer, Michael (Hg.): Europäische Öffentlichkeit und medialer Wandel. Eine transdisziplinäre Perspektive. Wiesbaden: VS Verlag für Sozialwissenschaften. S. 93-106

[23] Koopmans, Ruud (2002): Codebook for the analysis of political mobilization and communication in European public spheres. Codebook from the Project: The Transformation of Political Mobilization and Communication in European Public Spheres. 5th Framework Program of the European Commission. Europub.com

[24] Koopmans, Ruud/Statham, Paul (1999): Political Claims Analysis: Integrating Protest Event and Public Discourse Approaches. In: Mobilization 4(2): pp. 203-222

[25] Luhmann, Niklas (2004): Die Realität der Massenmedien. 3. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften

[26] Maier, Michaela, Jens Tenscher (2007) (eds.): Campaigning in Europe - Campaigning for Europe: Political Parties, Campaigns, Mass Media and the European Parliament Elections 2004.

[27] Marshall, Thomas H. (1977): Class, Citizenship and Social Development. Chicago; London: University Press.

[28] McDonald, Daniel G./Dimmick, John (2003): The Conceptualiziation and Measurement of Diversity. In: Communication Research Vol. 30, No. 1/2003. S. 60-79

[29] McQuail, Denis (2000): Mass communication theory. 4thed. London/Thousand Oaks/New Delhi: Sage

[30] McQuail, Denis (2005): McQuail's Mass Communication Theory. 5th edition. London/Thousand Oaks/New Delhi: Sage

[31] Münz, Rainer (2008): Migration, Labor Markets, and Integration of Migrants: An Overview for Europe, World Bank Report.

[32] Ogris, Günther (1995): Social Integration of Ethnic groups and Xenophobia. MA thesis, Essex University

[33] Pfetsch, Barbara/ Adam, Silke/Berkel, Barbara/Medrano, Juan Diez (2004): Integrated Report: The Voice of the Media in European Public Sphere: Comparative Analysis of Newspaper Editorials. Integrated Report WP 3, Project The Transformation of Political Mobilization and Communication in European Public Spheres. 5th Framework Program of the European Commission. Europub.com

[34] Stevenson, Nick (2001): „Culture and Citizenship: an Introduction". In: Stevenson, Nick (Hrsg.) Culture and Citizenship. London (u.a.): Sage, S.1 – 10.

[35] Van Cuilenburg, Jan (2000): On Measuring Media Competition and Media Diversity. Concepts, Theories and Methods. In: Picard, Robert G. (Ed.): Measuring Media Content, Quality, and Diversity. Approaches and Issues in Content Research. Turku: Turku School of Economics. pp. 51-84

[36] Förster, Jens (2007): Kleine Einführung in das Schubladendenken. Vom Nutzen und Nachteil des Vorurteils. München: Dt. Verl.-Anstalt

[37] Wilhelm Heitmeyer (2008) (ed.): Deutsche Zustände. Band 7, Frankfurt a. M.: edition suhrkamp

[38] Krippendorf, Klaus (2004): Content analysis. An introduction to its methodology. 2nd edition, London/Thousand Oaks/New Delhi: Sage

[39] Früh, Werner (2007): Inhaltsanalyse: Theorie und Praxis. 6., überarb. Aufl. Konstanz: UVK

[40] Merten, Klaus (1995): Inhaltsanalyse. Einführung in Theorie, Methode und Praxis. Opladen: Westdt. Verlag.

[41] Roberts, C.W. (2001): Content Analysis. In: Smelser, Neil J./Baltes, Paul B. (ed.): International Encyclopedia of the Social & Behavioral Sciences. Amsterdam/Paris/New York/Oxford/Shannon/Singapore/Tokyo: Elsevier

[42] Neuendorf, Kimberly A. (2007): The content analysis guidebook. Thousand Oaks, Calif., Sage

[43] Saurwein, Florian / Brantner, Cornelia / Dietrich, Astrid (2006): Europäisierung der österreichischen Öffentlichkeit: Mediale Aufmerksamkeit für EU-Politik und der veröffentlichte Diskurs über die EU-Erweiterung. Forschungsbericht im Rahmen des Forschungsprogramms "new orientations for democracy in Europe" (node) des Bundesministeriums für Bildung, Wissenschaft und Kultur (bm:bwk).

[44] Zick, Andreas/Wolf, Carina/Küpper, Beate/Davidov, Eldad/Schmidt, Peter/Heitmeyer, Wilhelm (2008). The syndrome of group-focused enmity: Theory and test. Journal of Social Issues, 64 (2), 363-383.

References from Semiotics

[45] Baldry, Anthony (2005). A Multimodal Approach to Text Studies in English: The role of MCA in multimodal concordancing and multimodal corpus linguistics. Campobasso: Palladino.

[46] Baldry, Anthony, Beltrami, Michele (2005). "The MCA Project: concepts and tools in multimodal corpus linguistics". In Maj Asplund Carlsson, Anne Løvland and Gun Malmgren (eds.), Multimodality: Text, culture and use: Proceedings of the Second International Conference on Multimodality. Kristiansand: Agder University College and Norwegian Academic Press, pp. 79-108.

[47] Baldry, Anthony O'Halloran, Kay (in press) Multimodal Corpus-Based Approaches to Website Analysis. London: Equinox.

[48] Baldry, Anthony, Thibault Paul J. (2006a). Multimodal Transcription and Text analysis. A multimedia toolkit and coursebook. London and New York: Equinox.

[49] Baldry, Anthony, Thibault, Paul, (2006b) "Multimodal corpus linguistics". In Geoff Thompson & Susan Hunston (eds.), System and Corpus: Exploring connections. London: Equinox, pp. 164-83, 2006.

[50] De Souza, C. S. (2005). The Semiotic Engineering of Human-Computer Interaction: Camb. Mass/London: Mit Press.

[51] Evert, Stefan. (2002-2004). The UCS Toolkit. Available online from http://www.collocations.de/software.html.

[52] Firth, John Rupert (1957). Papers in linguistics 1934-1951. London: Oxford University Press.

[53] Kress, Gunther, Van Leeuwen, Theo (2006[1996]). Reading Images: The grammar of visual design. London and New York: Routledge.

[54] Halliday, Michael (1978). Language as Social Semiotic: The social interpretation of language and meaning. London: Edward Arnold.

[55] Halliday, Michael (1989) "Part A". In Michael Halliday, Ruqaiya Hasan (eds.), Language, Context and Text: Aspects of language in a social-semiotic perspective. Oxford: Oxford University Press, pp.1-49.

[56] Hare, J. S., Lewis, P. H., Enser P. G. B., Sandom C. J. (2006a) "A Linear-Algebraic Technique with an Application in Semantic Image Retrieval". Image and Video Retrieval: 5th International Conference, CIVR 2006, Tempe, AZ, USA, LNCS 4071 .pp. 31-40. ISSN 0302-9743, July 2006.

[57] Marenzi, Ivana (in press) "Designer genres: social, interactional, technological and multimodal aspects of Web2". In Anthony Baldry and Elena Montagna (eds), in press, Interdisciplinary Perspectives on Multimodality: Theory and practice. Proceedings of the Third International Conference on Multimodality, Palladino, Campobasso.

[58] Scollon, Ron. (2004). (with Philip LeVine). "Multimodal discourse analysis as the confluence of discourse and technology". In Philip LeVine and Ron Scollon (eds.) Discourse and technology: Multimodal discourse analysis. Georgetown University Round Table on Languages and Linguistics: Washington, DC: Georgetown University Press.

[59] Scollon, Ron (2001). Mediated Discourse: The nexus of practice. London and New York: Routledge.

[60] Sinclair, John (1998) "Corpus evidence in language description", In Gerry Knowles, Tony McEnery, Stephen Fligelstone, Anne Wichman, (Eds.) Teaching and language corpora . Longman pp. 27-39.

References from Library and Information Science

[61] Broughton, V., & Lane, H. (2000). Classification schemes revisited: Applications to Web indexing and searching. Journal of Internet Cataloguing, 2(3/4), 143-155.

[62] Broughton, V. (2004). Faceted classification: a tool for subject access in the twenty first century. Signum, (8), 5-18.

[63] Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings; 58(1/2), 49-72.

[64] Campbell, D. J. (1957). Glossary to Dr. Ranganathan's paper. Proceedings of the International Study Conference on Classification for Information Retrieval. London: Aslib.

[65] Gopinath, M.A. (1986). Construction of Depth Version of Colon Classification. New Delhi: Wiley Eastern.

[66] La Barre, K. (2004a). Adventures in faceted classification: A brave new world or a world of confusion? In I. McIlwaine (Ed.), Advances in Knowledge Organization: Knowledge Organization and the global information society (Vol. 9. pp. 79-84). Wurzburg: Ergon Verlag.

[67] La Barre, K. (2004b). The art and science of classification: Phyllis Allen Richmond, 1921-1997. (2004). Library Trends, 52(4), 765-791.

[68] Mills, J. (1960). A modern outline of library classification. London: Chapman & Hall.

[69] Mills, J. (2004). Faceted classification and logical division in information retrieval. Library Trends, 52(3), 541-570.

[70] Ranganathan, S. R. (1937/1957/ 1967). Prolegomena to library science. New York: Asia Publishing.

[71] Ranganathan, S. R. (1945/1960/1962). Elements of library classification. New York: Asia Publishing.

[72] Ranganathan, S. R. (1951a). Philosophy of library classification. Copenhagen: Ejnar Munksgaard.

[73] Ranganathan, S. R. (1951b). Classification and communication. Delhi, India: University of Delhi.

[74] Vickery, B. C. (1951). The structure of a connective index. Journal of Documentation,6(3), 140-151.

[75] Vickery, B. C. (1959). CRG Bulletin 5 -Construction of a classification scheme for aeronautics Cranfield. Journal of Documentation, 15, 39-57.

[76] Vickery, B. C. (1960). Faceted classification: A guide to construction and use of special schemes. London: Aslib.

[77] Vickery, B. C. (1966). Faceted classification schemes. In S. Artandi (Ed.), Rutgers Series on Systems for the Intellectual Organization of Information (Vol. 5). New Brunswick,NJ: Graduate School of Library Science at Rutgers University.

[78] Bhattacharyya, G. (1979). POPSI: its fundamentals and procedure based on a general theory of subject indexing languages. Library Science with a Slant to Documentation, Vol. 16 No. 1, March, pp. 1-34.

[79] Reitz, J. (2004). Online Dictionary of Library and Information Science. Retrieved 6 May, 2006 from http://lu.com/odlis/

[80] Prasad, ARD and N. Guha. Expressing Faceted Subject indexing in SKOS/RDF. In International Conference of Semantic Web and Digital Libraries, Bangalore 21 23, February, 2007.

[81] Prasad, ARD and D. Madalli. Faceted Infrastructure for Semantic Digital Libraries. Library Review. Forth coming.

References for Knowledge Evolution

[82] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, Oren Etzioni: Open Information Extraction from the Web, IJCAI 2007

[83] Klaus Berberich, Srikanta Bedathur, Mauro Sozio, Gerhard Weikum: Bridging the Terminology Gap in Web Archive Search, WebDB 2009

[84] Philipp Cimiano, Johanna Völker: Text2Onto, NLDB 2005

[85] Alessandro Moschitti, Daniele Pighin, Roberto Basili: Tree Kernels for Semantic Role Labeling, Computational Linguistics 34(2), Special Issue on Semantic Role Labeling, 2008

[86] M. Palmer, P. Kingsbury, D. Gildea: The Proposition Bank: An Annotated Corpus of Semantic Roles, Computational Linguistics 31(1), 2005

[87] Bo Pang, Lillian Lee: Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval 2(1-2), 2008

[88] Marius Pasca: Towards temporal web search, SAC 2008

[89] Sunita Sarawagi: Information Extraction, Foundations and Trends in Databases 1(3), 2008

[90] Snowball and QXtract: Scalable Information Extraction over Large Document Collections, http://snowball.cs.columbia.edu/

[91] Fabian M. Suchanek, Georgiana Ifrim, Gerhard Weikum: Combining linguistic and statistical analysis to extract relations from web documents, KDD 2006

[92] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum: YAGO: A Large Ontology from Wikipedia and WordNet, Journal of Web Semantics 6(3): 203-217, 2008

[93] Fabian M. Suchanek, Mauro Sozio, Gerhard Weikum: SOFIE: a self-organizing framework for information extraction, WWW 2009

[94] TARSQI: Temporal Awareness and Reasoning Systems for Question Interpretation, http://www.timeml.org/site/tarsqi/index.html

[95] Nina Tahmasebi, Tereza Iofciu, Thomas Risse, Claudia Niederee, Wolf Siberski: Terminology Evolution in Web Archiving: Open Issues, IWAW 2008

[96] Kristina Toutanova, Aria Haghighi, Christopher D. Manning: A Global Joint Model for Semantic Role Labeling, Computational Linguistics 34(2), Special Issue on Semantic Role Labeling, 2008

[97] Tomasz Tylenda, Ralitsa Angelova, Srikanta Bedathur: Towards Time-aware Link Prediction in Evolving Social Networks, KDD-SNA 2009

[98] Gerhard Weikum, Gjergji Kasneci, Maya Ramanath, Fabian M. Suchanek: Database and information-retrieval methods for knowledge discovery, Communications of the ACM 52(4), 2009

[99] Feiyu Xu, Hans Uszkoreit, Hong Li: A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity, ACL 2007

[100] Qi Zhang, Fabian M. Suchanek, Lihua Yue, Gerhard Weikum: TOB: Timely Ontologies for Business Relations, WebDB 2008

[101] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, Ji-Rong Wen: StatSnowball: a statistical approach to extracting entity relationships, WWW 2009

Other References

[102] E. Agichtein and L. Gravano. Snowball: Extracting Relations from Large Plain Text Collections. In ICDL 2000.

[103] E. Agichtein and S. Sarawagi. Scalable Information Extraction and Integration. Tutorial, KDD 2006.

[104] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, 2009. Diversifying search results. In WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pages 5{14, New York, NY, USA, 2009. ACM.

[105] T. Arni, P. Clough, M. Sanderson, M. Grubinger. Overview of the imageclefphoto 2008 photographic retrieval task. In Retrieved 18-06-2009, http://www.clef-campaign.org/2008/working notes/ImageCLEFphoto2008-¯nal.pdf, 200

[106] R. Baeza-Yates, L. Calder´on-Benavides, C. Gonz´alez-Caro, 2006. The Intention Behind Web Queries. In Proceedings of String Processing and Information Retrieval (SPIRE 2006). Glasgow, Scotland, pp. 98-109.

[107] P. Beineke, T. Hastie, C. Manning and S. Vaithyanathan. An exploration of sentiment summarization. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2003.

[108] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, D. Jurafsky. Extracting opinion propositions and opinion holders using syntactic and lexical cues. Computing Attitude and Affect in Text: Theory and Applications, pp. 125-141, 2006.

[109] Cunningham. GATE, A General Architecture for Text Engineering. Computing and the Humanities, Vol. 36:223-254, 2002.

[110] R. Datta, D. Joshi, J. Li, and J. Z. Wang, 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2), pp. 1-60, 2008.

[111] K. Dave, S. Lawrence and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In WWW '03: Proceedings of the 12th international conference on World Wide Web, pages 519–528, New York, NY, USA, ACM, 2003.

[112] X. Ding, B. Liu and P. S. Yu. A holistic lexicon-based approach to opinion mining. In WSDM '08: Proceedings of the international conference on Web search and web data mining, pages 231–240, New York, NY, USA. ACM, 2008.

[113] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, pages 617–624, New York, NY, USA. ACM, 2005.

[114] O. Etzioni, M. J. Cafarella, D. Downey, Ana Maria Popescu, Tal Shaked, S. Soderland, D. S. Weld and A. Yates. Unsupervised named entity extraction from the Web: An experimental study. Artif. Intell. 165(1), 91-134, 2005.

[115] Fairspin, http://fairspin.org (last access: 2009/07/08)

[116] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst and A. C. König. BLEWS: Using Blogs to Provide Context for News Articles. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2008.

[117] F. Giunchiglia, 2006. Managing Diversity in Knowledge. M. Ali and R. Dapoigny (Eds.): IEA/AIE 2006, LNAI 4031, p. 1, 2006. Springer-Verlag Berlin Heidelberg. Slides at: http://www.disi.unitn.it/~fausto/knowdive.ppt

[118] S. Gollapudi, A. Sharma, 2009. An axiomatic approach for result diversification. In WWW '09: Proceedings of the 18th international conference on World Wide Web, pp. 381-390, New York, NY, USA, 2009. ACM.

[119] J. S. Hare and P. H. Lewis, 2005. On image retrieval using salient regions with vector-spaces and latent semantics. In CIVR, W. K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, eds., LNCS 3568, pp. 540-549, Springer, 2005.

[120] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In Proceedings of ACL 97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174–181, Madrid, ES, 1997.

[121] M. Hu and B. Liu. Mining and summarizing customer reviews. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177, New York, NY, USA. ACM, 2004.

[122] ImageCLEF, http://www.imageclef.org (last access: 2009/07/08)

[123] Youngho Kim, Yuchul Jung, and Sung-Hyon Myaeng. Identifying opinion holders in opinion text from online newspapers. In GRC '07: Proceedings of the 2007 IEEE Inter-national Conference on Granular Computing, page 699, Washington, DC, USA, 2007.

[124] Soo-Min Kim, Eduard Hovy, 2006. Extracting opinions, opinion holders, and topics ex-pressed in online news media text. In Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, pages 1-8. Association for Computational Linguistics, 2006.

[125] N. Koudas, S. Sarawagi and D. Srivastava. Record Linkage: Similarity Measures and Algorithms. Tutorial SIGMOD, 2006.

[126] Gunther Kress (ed.) 1998 Communication and culture. An Introduction. Third Edition. Kensington: University of New South Wales Press.

[127] B. Liu. Web Data Mining. Springer, 2007.

[128] B. Liu, M. Hu and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In WWW '05: Proceedings of the 14th international conference on World Wide Web, pages 342–351, New York, NY, USA. ACM, 2005.

[129] C. McKay, A. Swaminathan, H. Gou, and M. Wu, 2008. Image Acquisition Forensics: Forensic Analysis to Identify Imaging Source. In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing, ICASSP2008, Las Vegas, NV, 2008, pp. 1657-1660.

[130] Alessandro Moschitti, Kernel Methods, Syntax and Semantics for Relational Text Categorization. In proceeding of ACM 17th Conference on Information and Knowledge Management (CIKM). Napa Valley, California, 2008.

[131] Newssift, http://www.vewssift.com (last access: 2009/07/08)

[132] Pasca et al. Organizing and Searching the World Wide Web of Facts — Step One: the One Million Fact Extraction Challenge. In AAAI, 2006.

[133] A. M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. Association for Computational Linguistics. In HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 339-346, Morristown, NJ, USA, 2005.

[134] A. C. Popescu and H. Farid, 2004. Statistical tools for digital forensic. In Proc. 6th Int. Work. on Information Hiding, IH'04, Toronto, Canada, 2004, vol. 3200, pp. 128–147.

[135] Tuffield, M., Harris, S., Dupplaw, D. P., Chakravarthy, A., Brewster, C., Gibbins, N., O'Hara, K., Ciravegna, F., Sleeman, D., Wilks, Y. and Shadbolt, N. R., 2006. Image annotation with Photocopain. In: 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006) at WWW2006, pages 22-26, 2006, Edinburgh, UK.

[136] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst., 21(4):315–346, 2003.

[137] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing, pages 129–136, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[138] D. Zelenko, C. Aone, A. Richardella. Kernel Methods for Relation Extraction. Journal of Machine Learning Research 3: 1083-1106, 2003.

[139] Ciaramita, Massimiliano and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 594–602. Sydney, Australia, 2006.

[140] Collins, Michael. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In Proceedings of the 2002 EMNLP Conference, pp. 1–8. Philadelphia, United States, 2002.

[141] Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Schwartz, and Yoram Singer. Online Passive-Aggressive Algorithms. Journal of Machine Learning Research, 7 (mar), pp. 551–585, 2006.

[142] Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, and Xiang-Rui Wang and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, pp. 1871–1874, 2008.

[143] Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task, pp. 1–18. Boulder, United States, 2009.

[144] Johansson, Richard and Pierre Nugues. Dependency-based Syntactic–Semantic Analysis with PropBank and NomBank. In CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning, pp. 183–187. Manchester, United Kingdom, 2008.

[145] Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young and Ralph Grishman. The NomBank Project: An Interim Report. In HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation. Boston, United States, 2004.

[146] Palmer, Martha, Dan Gildea and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31 (1), pp. 71–106, 2005.

[147] Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning, pp. 159–177. Manchester, United Kingdom, 2008.

[148] Wiebe, Janyce, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39 (2-3), pp. 165–210, 2005.

[149] Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. Leila: Learning to extract information by linguistic analysis. In Second Workshop on Ontology Population (OLP2) at ACL/COLING, 2006.

[150] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Web Semantics, 2008.

[151] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. SOFIE: A Self- Organizing Framework for Information Extraction. In International World Wide Web conference (WWW 2009), New York, NY, USA, 2009. ACM Press.

[152] Y. Rui, T.S., Huang, and S.-F., Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," Journal of Visual Communication and Image Representation, 10(4):39-62, 1999.

[153] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, 22(12):1349–1380, 2000.

[154] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, 2008.

[155] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval," IEEE Trans. Circuits and Systems for Video Technology, 8(5):644–655, 1998.

[156] Thomas Sikora, "The MPEG-7 Visual Standard for Content Description - An Overview", IEEE Trans. Circuits and Systems for Video Technology, vol.11, n. 6, June 2001.

[157] J. Kennedy, R. C. Eberhart, and Y. Shi, Swarm Intelligence. San Francisco, CA: Morgan Kaufmann, 2001.

[158] R. C. Eberhart and Y. Shi, "Particle swarm optimization: developments, applications and resources," in Proceedings of Congr. Evolutionary Computation, vol. 1, 2001.

[159] K. Chandramouli and E. Izquierdo, "Image retrieval using particle swarm optimization", In Advances in Semantic Media Adaption and Personalization, pp 297, Auerbach Publications, 2009.

[160] M. Broilo, P. Rocca, and F.G.B. De Natale, "Content-Based Image Retrieval by a Semi-Supervised Particle Swarm Optimization", International Workshop on Multimedia Signal Processig 2008.

[161] J. Fridrich, "Digital image forensics", IEEE Signal Processing Magazine, 26(2), pp. 26-37, 2009

[162] H. Farid, "A Survey of Image Forgery Detection", IEEE Signal Processing Magazine, 26(2), pp. 16-25, 2009

[163] A. Swaminathan, M. Wu, K.J.R. Liu, "Digital Image Forensics via Intrinsic Fingerprints", IEEE Transactions on Information Forensics and Security, 3(1), pp. 101-117, 2008.

[164] S. Lyu, "Natural Image Statistics for Digital Image Forensics", Ph.D. Dissertation, Department of Computer Science, Dartmouth College, 2005.

[165] T. T. Ng, "Statistical and Geometric Methods for Passive-blind Image Forensics", PhD Dissertation, Graduate School of Arts and Sciences, Columbia University, 2007.

[166] C. McKay, A. Swaminathan, H. Gou, M. Wu, "Image acquisition forensics: forensic analysis to identify imaging source", IEEE International Conference on Acoustic, Speech, and Signal Processing, Las Vegas, NV, March 2008.

[167] Alin C. Popescu and Hany Farid , "Exposing digital forgeries by detecting traces of re-sampling", IEEE Transactions on Signal Processing, 53(2), pp. 758-767, 2005.

[168] Babak Mahdian and Stanislav Saic, "Blind authentication using periodic properties of interpolation", IEEE Transactions on Information Forensics and Security, 3(3), pp. 529-538, 2008.

[169] Matthias Kirchner, "Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue", in Proc. 10th ACM Workshop on Multimedia and Security, Oxford, UK, pp. 11-20, 2008.

[170] Zhigang Fan and Ricardo L. de Queiroz, "Identification of bitmap compression history: JPEG detection and quantizer estimation", IEEE Transactions on Image Processing, 12(2), pp. 230-235, 2003.

[171] Jan Lukàs and Jessica Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images", Proc. of DFRWS 2003, pp. 5-8, 2003.

[172] Weiqi Luo and Zhenhua Qu and Jiwu Huang and Guoping Qui, "A novel method for detecting cropped and recompressed image block", in Proc. of 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2007, pp. II-217--II-220, 2007.

[173] Hany Farid, "Exposing digital forgeries from JPEG ghosts", IEEE Transactions on Information Forensics and Security, 4(1), pp. 154-160. 2009.

[174] Jessica Fridrich and David Soukal and Jan Lukàs, "Detection of copy-move forgery in digital images", Proceedings of DFRWS 2003, 2003.

[175] Alin C. Popescu and Hany Farid, "Exposing digital forgeries by detecting duplicated image regions", in Technical Report TR2004-515, Department of Computer Science, Dartmouth College, 2004.

[176] Babak Mahdian and Stanislav Saic, "Detection of copy-move forgery using a method based on blur moment invariants", Forensic Science International,171(2—3), pp. 180-189, 2007.

[177] Sevinc Bayram and Husrev Taha Sencar and Nasir Memon, "An Efficient and Robust Method For Detecting Copy-Move Forgery", in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, 2009.

[178] Avcibas, I.; Bayram, S.; Memon, N.; Ramkumar, M.; Sankur, B., "A classifier design for detecting image manipulations", in Proc. of IEEE International Conference on Image Processing, ICIP '04, 4, pp. 2645-2648, 24-27 October 2004.

[179] Stamm, M.; Liu, K.J.R., "Blind forensics of contrast enhancement in digital images", in Proc. of IEEE International Conference on Image Processing, ICIP 2008, pp. 3112-3115, 12-15 October 2008.

[180] S. Lyu and H. Farid, "How Realistic is Photorealistic?", IEEE Trans. on Signal Processing, vol. 53(2), pp. 845-850, 2005.

[181] Y. Wang and P. Moulin. "On Discrimination between Photorealistic and Photographic Images and Computer Generated", Proc. of IEEE ICASSP, pp. 161-164, 2006.

[182] T.-T Ng, S. -F. Chang, J. Hsu, L. Xie, M. -P. Tsui, "Physics-Motivated Features for Distinguishing Photographic Images and Computer Graphics", ACM Multimedia, pp. 239-248, 2005.

[183] S. Dehnie, H.T. Sencar and N. Memon, "Identification of Computer Generated and Digital Camera Images for Digital Image Forensics", Proc. of IEEE ICIP, pp. 2313- 2316, 2006.

[184] E. Dirik, S. Bayram and N. Memon, "New Features to Identify Computer Generated Images", Proc. of IEEE ICIP, 2007.

[185] F. Pan, J. Chen and J. Huang, "Discriminating between photorealistic computer graphics and natural images using fractal geometry", Science in China Series F: Information Sciences, vol. 52, No. 2. , pp. 329-337, 2009.

[186] M.K. Johnson, and H. Farid, "Detecting Photographic Composites of People", 6th International Workshop on Digital Watermarking, Guangzhou, China, 2007.

[187] J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In ACM CIVR '08, pages 359–368. ACM, July 2008.

[188] J. Tang and P. Lewis. Using multiple segmentations for image auto-annotation. In ACM International Conference on Image and Video Retrieval (CIVR), 2007.

[189] J. S. Hare, P. Lewis, L. Gordon, and G. Hart. Mapsnapper: Engineering an efficient algorithm for matching images of maps from mobile phones. In T. Gevers, R. Jain, and S. Santini, editors,

Multimedia Content Access: Algorithms and Systems II, volume SPIE Vol. 6820. SPIE and IS&T, January 2008.

[190] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In In CVPR, pages 2161–2168, 2006.

[191] Tuffield, M., Harris, S., Dupplaw, D. P., Chakravarthy, A., Brewster, C., Gibbins, N., O'Hara, K., Ciravegna, F., Sleeman, D., Wilks, Y. and Shadbolt, N. R. (2006) Image annotation with Photocopain. In: First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006) at WWW2006, May 2006, Edinburgh, United Kingdom.

[192] Biber, D. (1993), 'The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings', Computers and the Humanities, 26: 331-345

[193] Biber, D. and Conrad S. (2009), 'Register, Genre, and Style', Cambridge Textbooks in Linguistics Amazon

[194] Halliday, M. A. K. and R. Hasan (1989), 'Language, context and text: Aspects of language in a social-semiotic perspective', Oxford: Oxford University Press.

[195] Biber, D. and Jones J. K. (2005), 'Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles', Corpus Linguistics and Linguistic Theory 1 - 2.

[196] Meyer zu Eissen, S. and Stein, B. (2004), 'Genre Classification of Web Pages.', KI 2004: Advances in Artificial Intelligence, 27th Annual German Conference on AI, KI 2004, Ulm, Germany, September 20-24, 2004.

[197] Boese, E. S. and Howe, A. E. (2005). 'Effects of web document evolution on genre classification.' In Proceedings of the 14th ACM international Conference on information and Knowledge Management (Bremen, Germany, October 31 - November 05, 2005). CIKM '05. ACM, New York, NY, 632-639. DOI= http://doi.acm.org/10.1145/1099554.1099715

[198] [Stephen R. Garner, WEKA: The Waikato Environment for Knowledge Analysis, In Proc. of the New Zealand Computer Science Research Students Conference, 1995, 57-64

[199] George Karyp, Technical Report No 02-017, University of Minnesota, Department of Computer Science,{CLUTO} - A Clustering Toolkit}, 2003

[200] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[201] Annotation graphs as a framework for multidimensional linguistic data analysis Steven Bird and Mark Liberman, Towards Standards and Tools for Discourse Tagging, Proceedings of the Workshop, pp 1-10. Association for Computational Linguistics, 1999.

[202] M. Halliday, 1978. Language as Social Semiotic: The social interpretation of language and meaning. London: Edward Arnold.

[203] J. Van Cuilenburg, 2000. On Measuring Media Competition and Media Diversity. Concepts, Theories and Methods. In: Picard, Robert G. (Ed.): Measuring Media Content, Quality, and Diversity. Approaches and Issues in Content Research. Turku: Turku School of Economics. pp. 51-84

[204] Paul J. Thibault 2004. Brain, mind, and the signifying body. An ecosocial semiotic theory Continuum: London and New York. Foreword by Michael Halliday, p. 50

[205] D. McDonald, J. G./Dimmick, 2003. The Conceptualization and Measurement of Diversity. In Communication Research Vol. 30, No. 1/2003. S. 60-79

[206] D. McQuail, 2000. Mass communication theory. 4thed. London/Thousand Oaks/New Delhi: Sage

[207] Rada Mihalcea. SEMCOR Semantically tagged corpus. Technical report, SMU Group for Natural Language Processing. 1998.

[208] F. Giunchiglia, P. Shvaiko, M. Yatskevich, 2006. Discovering missing background knowledge in ontology matching. In ECAI, pages 382–386, 2006.

[209] F. Giunchiglia, M. Marchese, I. Zaihrayeu, 2006. Encoding Classifications into Lightweight Ontologies. Journal of Data Semantics 8, pp. 57-81, 2006.

[210] F. Giunchiglia, I. Zaihrayeu, 2008. Lightweight ontologies. In S. LNCS, editor, Encyclopedia of Database Systems, 2008.

[211] J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In ACM CIVR '08, pages 359–368. ACM, July 2008.

[212] J. Tang and P. Lewis. Using multiple segmentations for image auto-annotation. In ACM International Conference on Image and Video Retrieval (CIVR), 2007.

[213] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In In CVPR, pages 2161–2168, 2006.

[214] J. S. Hare, P. Lewis, L. Gordon, and G. Hart. Mapsnapper: Engineering an efficient algorithm for matching images of maps from mobile phones. In T. Gevers, R. Jain, and S. Santini, editors, Multimedia Content Access: Algorithms and Systems II, v. SPIE Vol. 6820. SPIE and IS&T, 2008.

[215] F. Giunchiglia, B. Dutta, V. Maltese, 2009. Faceted Lightweight Ontologies. In "Conceptual Modeling: Foundations and Applications", Alex Borgida, Vinay Chaudhri, Paolo Giorgini, Eric Yu (Eds.) LNCS 5600 Springer.

[216] David Day, Chad McHenry, Robyn Kozierok, and Laurel Riek. Callisto: A Configurable Annotation Workbench. In Proceedings of the LREC Workshop on Multimodal Annotation Tools, pp. 2073–2076. Lisbon, Portugal, 2004.

[217] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39 (2-3), pp. 165–210, 2005.

[218] Taku Kudo and Yuji Matsumoto. *Fast* Methods for Kernel-based Text Analysis. In Proceedings of ACL 2003.

[219] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best Parsing and MaxEnt Discriminative Reranking. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005.

[220] Kristina Toutanova, Aria Haghighi, and Christopher Manning. Joint Learning Improves Semantic Role Labeling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005.

[221] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. The Penn Discourse Treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation. 2008.

[222] Alessandro Moschitti. Kernel Methods, Syntax and Semantics for Relational Text Categorization. In Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM). 2008.

[223] Vincenzo Maltese, Fausto Giunchiglia, Kerstin Denecke, Paul Lewis, Cornelia Wallner, Anthony Baldry, Devika Madalli. On the interdisciplinary foundations of diversity. In Proceedings of the first Living Web Workshop at ISWC 2009.

[224] Jerome Euzenat. Towards a principled approach to semantic interoperability. In the Proceedings of the IJCAI-01, Workshop on Ontologies and Information Sharing, Seattle, USA, August 4-5, 2001.

Appendix A: An example of Media Content Analysis

This is an example of a Codebook[28] designed for (mass) mediated content (here especially for mass media articles). Based on the general model of MCA, the final Codebook will be applicable to mass media, press releases, forums and blogs. The list of variables gives an insight into the possibilities. In other words, this Codebook is not intended to be exhaustive.

For each variable, the name is given, followed by the set of possible values (codes) and optional guidelines for coding. Missing values are given the conventional code 999. In addition, note that the codes are normally listed in an Excel-template or SPSS-template. The codes are the starting point for further analysis, in keeping with specific research questions (e.g. diversity of speakers, opinion, frames etc.). For further use in (applied) social science research producing such an SPSS-file through automatic rather than manual annotation is highly significant. For data interpretation one could ask for instance for the relation Actor-Issue-Time and its evolution over time in the mediated discourse.

First check, if the article is concerned with "integration" in our understanding by reading headline, sub-line, lead text and first 150 words of the article (if it ends in the middle of a sentence, the rest of the sentence has to be included). If yes, continue coding. *Note: Done if manually coded, due to restrictions in time and capacity. Might be extended when automatically coded.*	
VARIABLES AT THE ARTICLE-LEVEL	
1. Name of variable: AID	**Article-identification number**
Rules for coding:	Give each article an ID with the first position indicating the content source. (list to be defined, referring to selected sources, e.g. 1-0001 BBC, 2-0001 Euronews, 3-0001 Economist etc.)
Values (codes) and labels:	running count per content source
2. Name of variable: Year	**Year of publication**
Values (codes) and labels	Use format YYYY
3. Name of variable: Month	**Month of publication**
Values (codes) and labels	Use format MM
4. Day	**Day of publication**
Codes	Use format DD
5. Source	**Source of Article**
	1 = mass media website 2 = official website 3 = private website Etc.
6. Size	**Size of Article**
Codes	Counted words
7. Position	**Is the article reported on the front page?**
	1 = yes 2 = no
8. Headline	**Headline of Article**
Codes	String variable: Type complete headline
9. Subline	**Subline of Article**
Codes	String variable: Type complete subline
10. Genre	**Genre of Article**
Codes	1 = news 2 = interview 3 = background article 4 = opinion
Rules for coding	1 News is coded, if it is a day-to-day coverage of events. 2 Interview is coded, if an interview carried out by a journalist is printed 3 Background article is coded for reports, background analysis, features 4 Opinion is coded, if a journalist/guest author/columnist gives his/her opinion to certain topics.
12. Topic	**Main topic of the article**

[28] The example provided refers to [23], [1], [12], [13], [14], [5] and [15].

	100 = immigration in general, immigration policy 200 = economics (incl. economic integration process of EU/nonEU-countries) 300 = integration (excl. economic integration process) 400 = other (→ specify as a string variable) *Note: normally topic plus 1st subtopic and 2nd subtopic is coded*
	Coding of the topic is done by the headline, subline, lead, first 150 words of text
NOTE: more variables to come on article level, according to the final model of content analysis, which will be used in LK	
End of coding on article-level	
VARIABLES AT THE STATEMENT(CLAIM)-LEVEL	
13. CID	**Identification number of the statement/claim**
	Count within each article 1-99
14. AID	**Article identification number of the claim**
	Same number as on article level
15. Title	**Title – description of the claim**
	Brief description of the claim, including main actor, form, addressees, object actor, aim. String_variable
16. Date	**Date of the claim YYYY-MM-DD**
17. Country	**Country affiliation**
	Country where the claim was made, use international abbr.
18. SACT1/2/3	**First subject actor → see list below** 100 Governments 200 Parliament: 300 Judiciary 400 Police, security 500 Military 600 Central banks 700 Social security organizations 800 Other state executive agencies (EU, European, national) 900 Political parties 1000 Media and journalists 2000 Civil society *Note: list of possible groups of actors with subgroups, 101, 102 etc.*
18. SACT1-TYP (2/3)	**Type of first subject actor** 1 = organisation, institution (e.g. EU-Federation of employees) 2 = representative of unorganised collective (e.g. "worker X") 3 = anonymous representative (e.g. "a worker") 4 = anonymous spokesperson for an organization/institution (e.g. "a spokesperson of the EU-Federation of employees…") 5 = unorganised collective, not specified / individual person 6 = named spokesperson for and organization/institution (e.g. "Person X, President of EU-Federation of employees")
19. SACT1-Name (2/3)	**Full name of first actor**
20. SACT1-Scope (2/3)	**Scope of first actor** 1 = supranational 2 = EU 3 = other European country (non-EUmember) 4 = national 5 = regional 6 = local
21. FORM	**Form of action** 1 = political decisions (legislation, parliamentary vote, administrative decision, ruling, personnel decisions etc.) 2 = executive action (arrest, criminal investigations etc.) 3 = judicial action (lawsuit) 4 = verbal statements (press conference, press release, interview, public speech,

	survey presentation, publicity campaign etc.) 5 = meetings (party congress, election campaign meeting etc.) 6 = direct-democratic action (referendum, initiative in CH) 7 = petitioning 8 = protest (all types of: demonstrative, confrontational, violent)
Addressee	
22. ADD	Addressee of speaker 1: the actor to whom the speaker refers in his/her claim (e.g. The churches called on to the government to do sth. → "churches" are speaker1, "the government" is addressee). Code like SACT1 *Note: furthermore, a distinction between opponent/supported actor is normally made*
23. ADD-Typ	See SACT1-TYP
24. ADD-Name	See SACT1-Name
25. ADD-Scope	See SACT1-Scope
Aim	
26. Issue1 (2/3)	**Issue of the claim** 100 = immigration in general, immigration policy (note: and subgroups) 200 = economics (incl. economic integration process of EU/nonEU-countries) (note: and subgroups) 300 = integration (excl. economic integration process) (note: and subgroups) 400 = other (→ specify as string variable) (note: and subgroups)
27. Issue-Scope1 (2/3)	**Scope of first issue**
	code like SACT1-Scope
28. Issue-country1 (2/3)	**Country to which Issue refers**
	Code like COUNTRY
29. Issue-rating1 (2/3)	**Relation of issue position towards European integration**
	1 = negative 2 = ambivalent/neutral 3 = positive (e.g. a claim is negative, if it opposes integration of migrants in EU-Society/labour market. A claim is positive, if it supports the integration of migrants in EU-society/labour market.)
Object actors	
29. OBJ1 (2/3)	**Object actor**
	The actor, whose interests are/would be affected by the claimant's claim. (e.g. The churches called on to the government not to deport migrants from country X. → The government is the addressee, the migrants from country X are object actor.) Code like SACT1
30. OBJ1-Typ (2/3)	**See SACT1-TYP**
31. OBJ1-Name (2/3)	**See SACT1-Name**
32. OBJ1-Scope (2/3)	**See SACT1-Scope**
33. OBJ1-effect (2/3)	**Effect of the claim on the object actor**
	1 = negative 2 = ambivalent/neutral 3 = positive The effect of the claim on an object actor is negative, if he/she is being criticised. The effect of the claim on the object actor is positive, if he/she is being supported e.g. "The spokesman x of the EU-Commission X applauded the efforts of the NGO Z for their suggestions about legal protections for children", which would be positive. "The politician X accused company Y of violating competition law.", which would be negative.

Frames	
Frames are coded within claims. We distinguish four frames for EU-integration, identity frames, instrumental frames, historical frames, frames internal to EU-integration process. *Note: For example, we use an existing list of frames referring to EU-integration (Pfetsch et al 2004). There are several methods of how to generate such lists for empirical research (see e.g. Matthes/Kohring 2004), which will be discussed in the part SoA media content analysis in social science research)*	
34. Idframe	**Identity frames**
	1 = principles, norms, values (with subcategories: peace, social equality, responsibility etc.) 2 = institutional, constitutional (with subcategories: human rights, rule of law, democracy etc.) 3 = governance (with subcategories: transparency, accountability, corruption etc.)
35. Inframe	**Instrumental frames**
	1 = political (with subcategories: stability, security etc. 2 = economical (with subcategories: taxes, inflation, economic growth etc.)
36. Hiframe	**Historical frames** Division into centuries, historical re-enactment
37. Ownframe	**Frames internal to EU-integration process**
	1 = political integration 2 = economic integration 3 = social integration 4 = enlargement in general
38. FrameSum	Summary of frame, stringvariable
39. FrameVerb	Verbal reproduction of frame, stringvariable

List of codes for SACT (Subject actors)

100 Governments:

- European Union/Community
- European Commission, Individual Commissioners, DG
- Council
- Other European but non-EU institutions
- UN organizations
- Other supranational/intergovernmental institutions
- National government/ministries

200 Parliament:

- European Parliament/Parliament Committee/Parliamentary Political Group
- EU Advisory Committee (Committee of the Regions, Economic and Social Committee)
- Council of Europe (Parliamentary Assembly)
- UN General Assembly
- National parliaments

300 Judiciary

- European Court of Justice, Court of first instance
- Non-EU European judiciary (European Court of Human Rights et al)
- United Nations judiciary
- Other supranational judiciaries
- National judiciaries

400 police, security

- EU security institutions (Europol et al)
- European but non-EU police/security institutions, co-operations
- UN police/security institutions
- Other supranational police/security institutions (Interpol et al)
- National police/security institutions

500 military

- EU military institutions/co-operation
- European but non-EU military institutions/co-operations

- UN military (blue helmet et al)
- Other supranational military bodies (Nato et al)
- National military bodies

600 central banks

- EU financial institutions, EU monetary institutions
- UN central bank
- Other supranational financial institution (IMF, WTO et al)

700 social security organizations
800 other state executive agencies (EU, European, national)

- E.g. European Monitoring Centre on Racism and Xenophobia, European Environment Agency etc.

900 political parties

- European political parties
- Supranational parties
- National parties

1000 media and journalists
2000 civil society

In the following we provide a small example of an article, the corresponding Codebook and an example of claim selection.

ARTICLE

Query: BBC (google: <integration "labour force" site:bbc.co.uk>)

Synopsis: In Germany's 2002 general election campaign immigration – despite claims to the contrary -- became once more an issue. This article talks about Stoiber's turn right and how Germany's messed up integration and citizenship laws hinder voting of "integrated" immigrants.

Original article:
Wednesday, 18 September, 2002, 11:58 GMT 12:58 UK
Immigration enters Germany's election
Clare Murphy,
BBC News Online, Berlin

Edmund Stoiber's Bavarian government once deported a German-born juvenile delinquent of Turkish parents "back" to Turkey - a country the young criminal had never set foot in.

In the last election battle waged by Mr Stoiber's conservative CSU party in 1998, immigration played a starring role. Posters plastered around the provinces advised voters not to back the CSU if they wanted more foreigners in the country.

This time the Bavarian politician, who is leading the country's conservative coalition into Sunday's general election, appeared to have dropped the topic from the campaign agenda in an attempt to moderate his hard-right image and woo the voters of the centre.

Until this week. Having lost its slight edge over Chancellor Gerhard Schroeder's Social Democrats in the opinion polls in this crucial, final stage, Mr Stoiber's coalition is clearly rattled.

Immigration - or the need to curb it - has reclaimed its place on the conservative to-do list.

"They are desperate to latch on to it now," says Halil Yildirim, sitting outside his furniture stall on the vibrant Bergmannstrasse in Berlin's Kreuzberg district, home to 150,000 people of Turkish origin - Germany's largest immigrant group.

"We do need some curbs on immigration - it will make life easier for the foreigners who have lived here for years. But it's totally wrong to make it an election issue."

Situation vacant

The Social Democrat-led government of Chancellor Gerhard Schroeder has looked to what it describes as a "controlled" immigration policy as a means of filling thousands of both skilled and unskilled vacancies in the German labour force.

Read Clare Murphy's campaign reports

Despite the country's doggedly high unemployment rates, hospitals and the catering industry in particular lack low-grade workers, the metal industry also has shortages, while professional posts lie open in firms across the

country.

Germany's declining birth rate also means that the country, like its European counterparts, desperately needs new recruits to work and contribute to its pay-as-you-go pension system.

To this end, a law was passed in May which made it easier for employers to look for labour abroad, a move broadly welcomed by industry.

Unconvinced

But Mr Stoiber says he is not having any of it. "When we've got more than four million jobless it is simply irresponsible to open the labour market to everyone," he told German voters this week, vowing to repeal the law if his conservatives are elected on Sunday.

The conservatives' priority: the integration of the seven million foreigners who do live in Germany, including sending them on compulsory language courses. Germany should not develop into a multi-cultural society, Mr Stoiber's interior affairs expert declared.

And their words were welcomed by some of those on the grimy streets of Kreuzberg today.

"I'm glad someone has finally brought it up," says Beate Bose, one of the district's few blonde residents.

"We can't just pretend that it's not a problem to have more foreigners who are prepared to work for low wages coming into Germany when we've got such a big problem of unemployment."

High-risk strategy

Mr Stoiber will be hoping that the immigrant gambit will play to concerns such as these, and push some of those millions of floating voters to turn out in support of the conservatives on Sunday.

But it is a gamble. The Bavarian leader has done his utmost to transform his image, and he runs the risk of alienating those voters who had started to sign up to the idea that Mr Stoiber was indeed a man of the centre - as well as those industries crying out for foreign labour.

One thing seems certain, though. Many of Kreuzberg's Turkish residents will not be turning out to vote in Sunday's election.

Despite a new law passed by Mr Schroeder's government which changed Germany's archaic citizenship laws and made it easier for foreigners to become German nationals with voting rights, the number of Turks applying for citizenship has fallen since the law came into force.

"It is a law which in fact discriminates against Turks," says Eren Unsal, spokeswoman for Berlin's Turkish community. "Under the new rules, we have to drop Turkish nationality if we apply for German citizenship, and many Turks are not prepared to do that."

"But I hope those who can do vote. The country's miserable forced integration and immigration policies will only get worse under Mr Stoiber."

(http://news.bbc.co.uk/2/low/europe/2265652.stm)

CODEBOOK: VARIABLES ON ARTICLE-LEVEL		
1. Name of variable: AID	**Article-identification number**	**1**
2. Name of variable: Year	**Year of publishing**	**2002**
3. Name of variable: Month	**Month of publishing**	**09**
4. Day	**Day of publishing**	**18**
5. Source	**Source of Article**	**1**
6. Size	**Size of Article**	**728**
7. Position	**Is the article reported on front page?**	**2**
8. Headline	**Headline of Article**	**Immigration enters Germany's election**
9. Subline	**Subline of Article**	**999**
10. Genre	**Genre of Article**	**3**
12. Topic	**Main topic of the article**	**1**
End of coding on article-level		

SELECTION OF CLAIMS
Main claim: *Quote* But Mr Stoiber says he is not having any of it. "When we've got more than four million jobless it is simply irresponsible to open the labour market to everyone," he told German voters this week, vowing to repeal the law if his conservatives are elected on Sunday.

The conservatives' priority: the integration of the seven million foreigners who do live in Germany, including sending them on compulsory language courses. Germany should not develop into a multi-cultural society, Mr Stoiber's interior affairs expert declared. *Quote*

Claim 1, Speaker 1, Issue 1, "When we've got more than four million jobless it is simply irresponsible to open the labour market to everyone," he told German voters this week
OBJ1: to open the labour market to everyone

Claim 1, Speaker 1, Issue 2: he told German voters this week, vowing to repeal the law if his conservatives are elected on Sunday.
ADD: German voters
OBJ1: German voters

Claim 1, Speaker 2: The conservatives' priority: the integration of the seven million foreigners who do live in Germany, including sending them on compulsory language courses. Germany should not develop into a multi-cultural society, Mr Stoiber's interior affairs expert declared.
Issue 1: The conservatives' priority: the integration of the seven million foreigners who live in Germany
OBJ1: foreigners

Issue 2: including sending them on compulsory language courses
OBJ1: foreigners

Issue 3: Germany should not develop into a multi-cultural society
ADD: Germany
OBJ1: Germany

Claim 2: *Quote* "They are desperate to latch on to it now," says Halil Yildirim, sitting outside his furniture stall on the vibrant Bergmannstrasse in Berlin's Kreuzberg district, home to 150,000 people of Turkish origin - Germany's largest immigrant group.
"We do need some curbs on immigration - it will make life easier for the foreigners who have lived here for years. But it's totally wrong to make it an election issue." *Quote*

Claim 3: *Quote* To this end, a law was passed in May which made it easier for employers to look for labour abroad, a move broadly welcomed by industry. *Quote*

Claim 4: *Quote* "I'm glad someone has finally brought it up," says Beate Bose, one of the district's few blonde residents.
"We can't just pretend that it's not a problem to have more foreigners who are prepared to work for low wages coming into Germany when we've got such a big problem of unemployment." *Quote*

Claim 5: *Quote* "It is a law which in fact discriminates against Turks," says Eren Unsal, spokeswoman for Berlin's Turkish community. "Under the new rules, we have to drop Turkish nationality if we apply for German citizenship, and many Turks are not prepared to do that."
"But I hope those who can do vote. The country's miserable forced integration and immigration policies will only get worse under Mr Stoiber."

Appendix B: Questions which can be answered by the MCA Web Browser

General characteristics

	Questions
01	Is the text a linguistic unit, a visual unit or a multimodal unit?
02	What dynamic elements are presented in the text. For example, a printed text is often assumed to be static in nature; however, many printed texts are dynamic, containing changes vis-à-vis the participants, circumstances and processes collapsed into a single image (e.g. a printed cartoon).

Experiential metafunctions (participants, processes & circumstances)
The Participants

03	Who/what are the participants in the text?
04	Who/what are the main participants?
05	Are they human, animal or inanimate?
06	Which participants are active?
07	Which participants are passive?
08	What ties exist between the participants?
09	How do they interact?
10	What activities are represented in the text?
11	What activities link up the participants?
12	What actions do the participants undertake?
13	Who/what does what to whom/what?
14	What relationship exists between the participants, the circumstances and the actions/activities depicted? For example: are the participants active, passive or reactive?
15	What changes are there in facial expression?
16	What can we learn from the participants' gestures?
17	Do the participants look at each other?
18	What physical changes occur? For example, do the participants bend over or lie down? … hide or cover themselves up? … protect themselves? … become temporarily deformed?
19	What mental changes occur?
20	What changes occur in emotional states?
21	What changes occur in behaviour? For example: do the participants start: laughing, crying or become angry?
22	What attitudes does the text suggest?

Processes: micro-processes

23	What material, verbal, mental, relational, existential and behavioural processes are represented?
24	What visual processes are involved (merging, separating, rotating, spinning?)
25	What metatextual processes are involved (captioning, labelling, naming)?

Processes: macro-processes

26	Is a story told?
27	Who tells the story?
28	What are the characteristics of the story?
29	Have you read/seen the first part of the story?
30	How do you think the story is likely to end?
31	Is the story told by one or more narrators?

Circumstances

32	What location or locations does the text represent?
33	What temporal scales are involved?
34	In what way are the participants shown to interact with physical space?
35	What contrasts exist in the spatial circumstances? For example, are the participants and processes partly in an internal environment and partly in an external environment?
36	In what way are the participants shown to interact with time?

37	What contrasts exist in the temporal circumstances? For example, are the participants and processes partly in one time and partly in another time?

Interpersonal metafunction (relationship between the text and the reader/viewer and within the text)

38	How does the text invite the reader/viewer to evaluate the actions represented in the text?
39	How does the text invite the reader/viewer to evaluate the events represented in the text?
40	Does the text follow a typical linear left-right top-down reading path or does it follow a more complex one? For example, are you invited to cluster-hop in a zig-zag fashion? If so why?
41	How does the reader negotiate the meanings made in the text? For example, are there elements that could be interpreted in different ways by different readers?
42	What interpretations are linked to the viewer's particular cultural background?
43	In what ways are the readers requested to become participants in the text?
44	How do distance, gaze and perspective contribute to creating a relationship between the participants in the text and the reader?
45	From what angle are the participants viewed? For example: front, side, rear …
46	How can the reader infer the narrative sequence in the text?
47	How is the reader led to adopt a particular interpretative stance on the text? For example, are the visual images sensuous, naturalistic, abstract or hyper-real?
48	How is the attention of the reader captured? For instance, are key-words used?
49	What expectations does the reader/viewer have about specific instances of the genre to which the text relates?
50	How does the reader read the text? What types of reading paths exist in this text ? For example, does the reader jump from running text to diagrams?
51	What ties exist between the participants? How are they represented in the text?

Special questions on identities that underpin and affect issues of bias and opinion

52	How far is identity defined by social roles?
53	How far do these roles overlap?
54	How is identity constructed?
55	How does the viewer interpret the identities established in the films in this corpus?
56	What is the politics of the notion of identity?
57	What roles do the body and physical interaction have in shaping identity?
58	What aspects of our identities are shaped by desire, pain and fear?
59	What kind of identities moves our society in new directions?
60	What identities are temporary and which are more permanent?
61	Do we construct identities only from fixed semiotic options provided by our culture and its constraints?
62	What role do transgressive identities play in social and cultural change and over what timescales?
63	What is the relationship between children and adults in terms of changing identity?
64	How are identities in this corpus shaped in terms of interaction with diverse members of our communities: elders/juniors; same gender/non-same gender, same class/non-same class?
65	What strengths and weaknesses are associated with identity?
66	How is identity constructed in terms of nationality, occupation, sexuality, cultural disposition, religion?
67	How is identity related to language?
68	How is language used to reinforce institutional stereotypes?
69	What is the relationship between the individual and the state?
70	How does the state use stereotypes to control us?
71	How has globalization affected identity? Is identity independent of globalization?

Textual/compositional metafunction (information management, resources, genres)
Macrolevel Resources; Cluster types/primary genres: e.g. speech & thought bubbles, logos slogans, slogos, tables, diagrams, graphs

72	What type of text is this in terms of its use of textual hierarchies?
73	What processes of recontextualisation exist?
74	How are textual resources used to represent the given/new and ideal/real dichotomies?
75	What primary genres (mini-genres) are there?
76	Is there more than one genre?
77	Is the text in question a scientific text or is the text a cartoon? Regardless of the genre, what is the management in text between given and new information and between the expected and the unexpected?
78	In what ways do visual modalities contribute to overlappings and mergings in the various experiential processes represented in the text? For example, does the page contain a graph representing changes over time in relation to inflation which shows unexpected peaks or troughs caused by major political events?

Micro level resources (i.e. below the level of subclusters)

79	What use is made of framing?
80	What visual and linguistic resources are expected/unexpected in the text?
81	What use is made of lines to build up higher scale objects e.g. clusters representing participants?
82	What use of colour is made? What is the significance of colour?
83	Do particular colours stand out?
84	What meanings are made by colour changes?
85	Are there any recognisable shapes: circles, rectangles, triangles, step-like structures?
86	What use of language is made?
87	Is language used to create a running text? If not why not and how?
88	Are the visual or linguistic resources ellipted?
89	Are the visual or linguistic resources convergent or divergent?
90	What use is made of dress, hairstyle, clothes, perfumes and other accoutrements?
91	What representation is given of the human body?
92	Can sound be a resource in printed texts? For instance, is onomatopoeia used in the text?
93	How is the resource integration principle used in the text? With what functions?
94	How is the meaning-compression principle used in the text? With what functions?
95	What resources are used to suggest attitudes?
96	What intertexts exist i.e. are you led to think about other texts? Which? Why are you led to think about them?
97	Are there links to the context of culture?
98	What resources are most salient? Why did you pick out these resources? What is the significance of these resources?
99	How are the text's various parts linked up?
100	What changes occur in resource patternings?
101	What textual units exist? For example, what clusters are there? Are there subclusters, clusters and superclusters?
102	What clusters function at a metatextual level, i.e. as captions, callouts, labels, overlays, subtitles?

Appendix C: Example of the application of MCA Web Browser

The goal of this appendix is to illustrate the hierarchy of semiotic units found in Web pages, running from the lowest to the highest level, in relation to the **MCA Web Browser** which uses manual and semi-automatic techniques and corpus-based annotation and retrieval procedures to identify the hierarchy of semiotic structures through which opinion and bias are typically enacted. The annotation procedures are based on the hypothesis that **multimodal meaning-making units** exist in Web pages that go beyond traditional categories such as written text and photos. An example is an online journal, such as the *Economist*, with its typical linkage between RUNNING TEXT, PHOTO and CAPTION, each of which are co-contextualising subparts of a single whole. Eye tracking tools provide empirical support for the existence of, and priority given to, these higher units. Recurring linkages, i.e. MULTIMODAL COLLOCATIONS, can be detected by the *MCA* system's corpus-based query searching tools.

Like *automatic extraction of contextual information* and *automatic image term annotation*, multimodal annotation is also in its infancy. However, multimodal annotation supports both issues. It addresses the key issue of *contextualisation* by defining it as a process through which integrations of visual, verbal and spatial resources – and not just surrounding words or *co-text* (see *Collocation* in Appendix D) – function at different *Rank* levels and mutually provide contextual information in Web pages, whence the notion of *Multimodal collocations* rather than *Language-only collocations* (see Appendix D). The annotation methodology used is itself multimodal. In other words, the corpus of annotations thus produced will, at a later stage in the project, be searchable and capable of identifying recurrent (inter)semiotic structures associated with the expression of opinion and bias. The corpus will have a pivotal role: it will assist in the progression from manual to automatic procedures (and vice-versa) and will constitute a means through which the various partners can collaborate at various stages in the project. Thus, it will help the project to proceed in an incremental, step-like manner. In particular, it is expected that the approach will be able to account for thematic expansions (viewed as the unfolding of meaning-making processes at the higher *Pagelet* and *SuperCluster* levels) which, as suggested in Fig. 13 and Fig. 14, form the basis of the recurrent and periodically-arranged patterning that typifies today's Web pages. We also expect to be able to carry out some form of automatic annotation for example in relation to the automatic capturing of written text found on a Web page.

Accessing *MCA Web Browser* (Fig. 20) it is possible to analyse the content of the first part of the "Special Report" as a *SuperCluster*, within a typical *Centre-Right Pagelet* (see Fig. 21 and Fig. 22).

Fig. 20 – The MCA Web Browser 2.0 home page

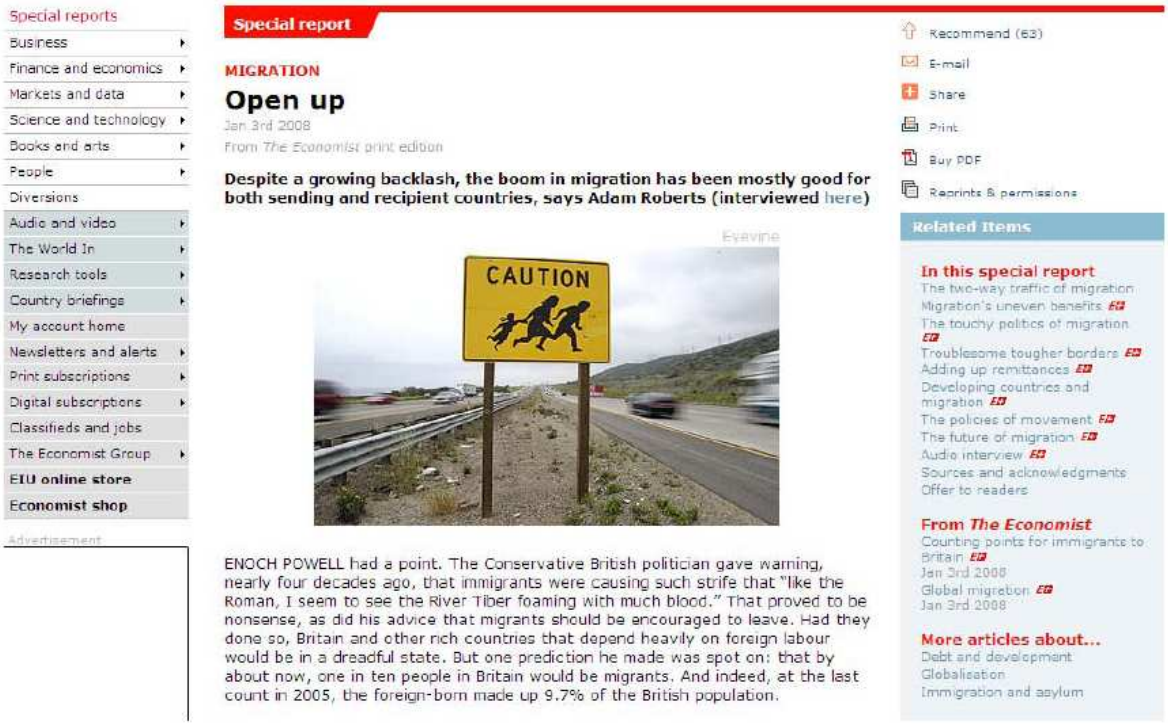

Fig. 21 - An example of pagey Web page (taken from the economist.com)

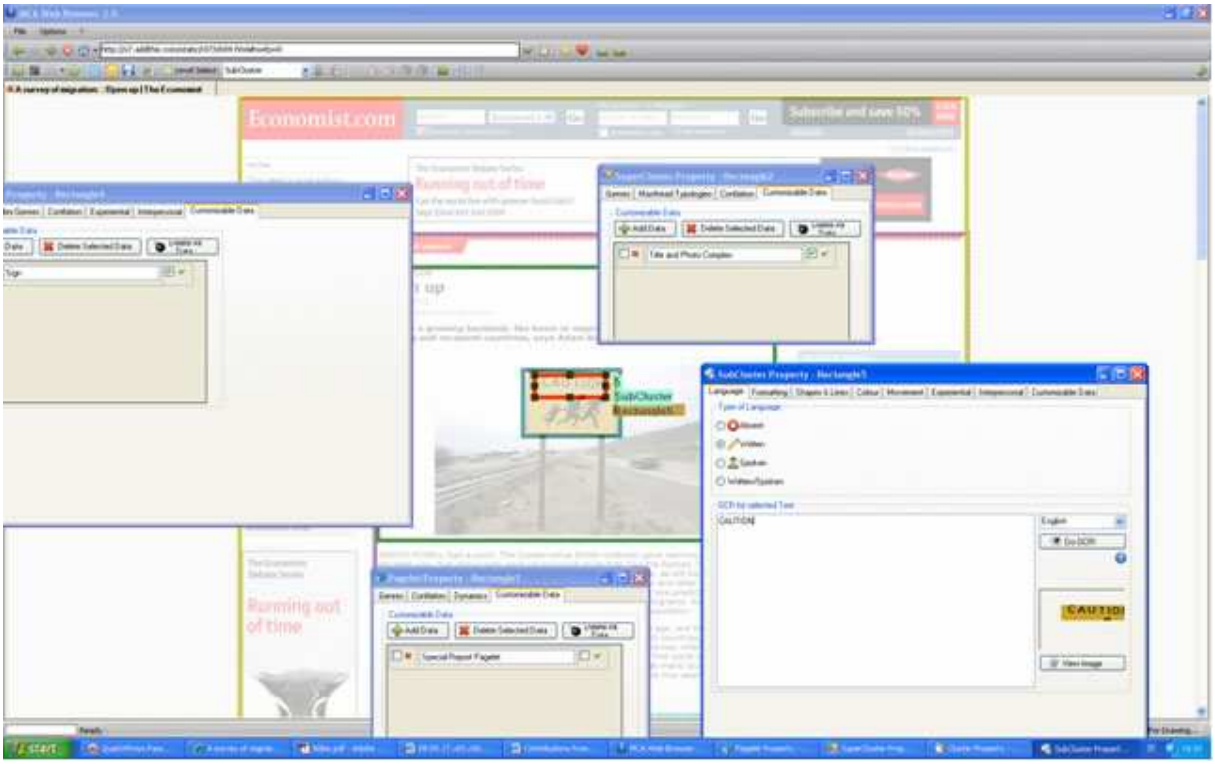

Fig. 22 – Multimodal annotation of A WEB PAGE from the ECONOMIST

An eye-tracking heat map for a part of this Web page (an aggregate of two users reading of the text, see Fig. 23) lends support for the overall annotation analysis. It suggests a primary focus on the centre-right *Pagelet* (i.e. after rapid scanning of the *Top Bar* and *Indexing Left Panel*) followed by a focus on the photo and the associated *Running Text*. The heat map suggests the *Photo* is analysed in terms of three subparts (*SubClusters* 19, 20, 21) and that the written text is also similarly subdivided (but analyzable in terms of well-established language-based micro annotations). We plan to use comparison of annotations by different users, corpus retrieval and eye-tracking techniques to validate the approach and to detect typical patterns of form and distribution of *Text Objects* in Websites.

Fig. 23 – A heat map generated by users looking at the top part of a Economist Webpage

Appendix D: Glossary of the terms used in Semiotics

COLLOCATION: Word-based collocations and multimodal collocations

Word-based collocations

Collocations are semiotic units consisting of words which frequently co-occur. Prefabricated expressions such as *naked eye* – a well-known construction first discussed by Sinclair [59] – and *blind eye*, which forms the core of the longer and more variable constructions as *turn a blind eye to something,* are examples of word-based collocations. How many times is the word order "white and black" found in English as compared with the normal collocation "black and white"? Armed with knowledge about collocations (and other textual phenomena) it is possible to carry out forensic explorations to establish whether written texts have been manipulated. This is because readers of texts intuitively 'tell [the meaning of] a word by the company it keeps' [52]. However, this intuitive capacity has much increased with the advent of query tools and Web corpora (e.g. *www.**webcorp**.org.uk/).* Extraction of collocations from language corpora by query-based tools is based on candidate words (such as *eye*) and the identification of their *co-text* (i.e. the words immediately to the right and left of the target words). Many query tools implement statistical measures based on word co-occurrence [51]. Thus, like a *sentence* or a *clause*, a *collocation* is ultimately a way of identifying rules about the way *texts* are put together.

Visual or multimodal collocations

"Smile and the whole world smiles with you" says an old saying. Not surprisingly, as Fig. 24 shows, "doing things with a smile" is a frequent pattern in Websites that often (but not always) correlates with positive assessments in accompanying *written text*. We need to find ways of pinning down this correlation and to do this we need to define the functions and status of smiles (or for that matter any other visual phenomenon). Understanding their hierarchical status in texts and photos (see *Scalar models* below) is significant in this respect: apart from exceptional cases (e.g. *Alice in Wonderland* where the Cheshire cat disappeared leaving *only* a grin) smiles will always be *secondary* items/events normally associated with socially more significant things: "doing things *(main activity)...* with a smile *(secondary supporting item/event)*".

Fig. 24 – Polish women selling sausages at the Borough Market in London (from the online version of the International Herald Tribune)

Visual or multimodal collocation is a term used to indicate those secondary items/events which do not have participant status, but which function to specify/contextualise either the role of the participant or the activity which he or she is performing. Visual collocations also indicate relationships, some recurrent, between these secondary items. In some texts, their use borders on the

stereotypical/ironical/sensational, insofar as each image (e.g. *photo* or *drawing*) is characterised by a number of such objects which function to index some aspect of the participant, his/her role, or the socially relevant location in which the depicted scene takes place. People walking with difficulty through the *snow* in the Arctic is not newsworthy; the same event in Singapore is. The *visual collocation* here is the relationship between snow and *other* secondary items such as buildings in Singapore that 'explains' and 'comments on' the main activity of walking with difficulty as exceptional.

Annotation with *MCA Web Browser* formalises these relationships on a par with the co-occurrence methods (i.e. word collocations) used with language. Thus the notion of collocation as used in Firthian and neo-Firthian approaches to language (see [59]) can be adapted to suggest the ways in which, for example, given objects, ways of dressing, occupational roles and institutional locations exhibit typical patterns of distribution in a visual field. The collocation serves to index the relevant situation or situation-type. For example, the holding of objects or tools can be indicative of a particular social status as Fig. 25 suggests. In the current state of research we do not know at what level or levels visual collocations typically occur in Websites (see *Scalar models of genres*) and to make progress we need to acquire this knowledge.

Fig. 25 – (on the left) Identifying an immigrant (from collectionscanada.gc.ca) ; (on the right) Identifying an immigrant as a protester (from usatoday.com)

TEXT AND DISCOURSE: Language-only texts, multimodal texts, mediated discourse

What is a text? What is a multimodal text? What is discourse? What is mediated discourse? Within semiotics and, in particular, within social semiotics, definitions of these terms are likely to differ. In many accounts *discourse* and *text* are virtually synonymous and refer to language in use (the view taken in this summary).

Language-only texts

Texts are defined following Halliday [55] as meaning-making events whose functions are determined by their use in particular social contexts. "Any living language that is playing some part in a context of situation, we shall call a text. It may be either spoken or written, or indeed in any other medium of expression that we like to think of" ([55]: 10). Examples include telephone calls and service encounters (spoken), letters (written), emails (which exhibit features of both written and spoken discourse) and Websites (which by definition are an example of multimedia). In this view, language is a basic low-level semiotic resource which is *used* in such a way as to carry out specific functions through typical *grammatical* and *textual* patterns functioning at a higher level (see *Scalar models of genres*). Language is thus *not* in itself text, but rather a resource with which to make texts. For example, *collocations* (see above) may be considered as low-level textual units; they are sub-clausal, prefabricated combinations of words. They are a *use/function* of the basic resource, language, and have the status of a *semiotic/textual unit*.

Multimodal texts

Multimodal texts combine more than one type of semiotic resource; they include semiotic resources such as colour, movement, gesture, gaze, space, time, shapes, lines, dots, shading, and spatial disposition; not all *multimodal texts* necessarily include language – for example, many road, railway, airport and other signs, many of which recur in Websites, do not. In a sense, all texts are multimodal, since language is itself multimodal, combining, for example in spoken forms, the resources of intonation, voice quality, breath control, rate of speaking, hesitations and pauses and, in written forms, (at the very least) the resources of font and punctuation. With the rise of multimedia technologies, the multimodal nature of texts is becoming increasingly more pronounced.

Mediated discourse

This viewpoint on discourse asks the questions "Which is more important in discourse: the material support for action and meaning *or* the action and meaning itself?" and "What is the difference between meaning and action?" It answers these questions by assuming that − *prior to analysis* − they are indistinguishable. "We use the phrase 'mediated action' to highlight the unresolvable dialectic between action and the material means which mediates all social action" ([59]:3). The *material* and the *semiotic* dimensions are thus the two sides of the same textual coin which are fully integrated into the one overall contextualising activity. Is a toy that a child is playing with merely a support for meaningful activity or is it action and meaning itself? By assuming that any object − a toy, a Website, a book, a car, a coin, a tattoo, a painting, a photo, a building − is *both* a material object *and* a semiotic object with a potential to instantiate meanings and actions relating to the present, past or future, it will be seen that this viewpoint, already important for written discourse, becomes even more significant with multimodal meaning making ([48]: 175) because of the greater analytical difficulty of separating the material and semiotic dimensions. Not surprisingly, given this challenge, many advocates of mediated discourse [54] carry out their theoretical investigations with reference to multimodality.

GENRE: Language-based genres and multimodal genres

Language based genres and multimodal genres

The term *genre* is used to indicate text types i.e. formal and functional recurrences which make it possible to distinguish a group of texts from other groups. For the purposes of this report, the terms *Genre, Discourse Genre, Text Genre, Multimodal Genre* are being used to refer to the forms and functions that groups of texts display. For example, the expressions "yellow pages", "web pages" and "teletext pages" indicate three text genres whose textual forms and social use are such to identify distinct genres.

Scalar models of genres

The current state of description of Websites and Web pages in terms of text function and text types is still limited. What, for example, does a *Masthead* do and how is it typically constituted in terms of semiotic form? What role does it play in a Web page and a Website? We do not know the answers to these questions. However, one conceptual tool that helps when describing the relationships between Website components and their functions, is a scalar model, a hypothesis about the semiotic organisation of Websites in terms of micro and macro levels.

	Highest → ← Lowest					
Text Levels	**Page**	**Pagelet**	**SuperCluster**	**Cluster**	**SubCluster**	**(Resource)**
Genre Levels	**Macro-genre**	**Hypergenre**	**Genre**	**Mini-genre**	**Genrelet**	**(Resource)**
	Thematic systems where abstract meanings are made			**Basic user interaction and interpretation**	**Basic resources: language, colour, gaze, gesture, movement, forms (lines and shapes), spatial disposition etc.**	

Fig. 26 – Meaning-making units in a Web page

This hypothesis (see Figure 26) distinguishes between a single *instance* of a meaning-making unit (i.e. *Text level*) and the possibility of classifying single instances into recurrent *types* or *genre* levels. In this view − reading the top row from right to left − a *Web Page* is a self-contained a multimodal unit of discourse quite unlike the archetype page in a book or newspaper printed, say, in the 19th century which had little functional autonomy vis-à-vis other pages and which was nothing more than a convenient division for printing. A modern Web page is, instead, a macro-unit whose *functional autonomy* is defined in terms of the subordinate functional levels it contains and whose complexity leads us to posit at least 5 levels of meaning making.

In order to function, a Web page mixes together basic meaning-making resources such as language, colour, lines and shapes at the very lowest level, the *SubCluster* level. These basic subparts make up, for example, the "Go" button in a *Search Engine*. The "Go" subpart is then associated to other subparts (such as the *Word type-in* box) to form a *Cluster* (in this case realising a *Search Engine*).

By applying these levels we can begin to identify the recurrent hierarchical patterns that distinguish a *Home Page* from a *Shopping Cart Page* and the many other types of Web pages which exist but for which, more often than not, names have not yet been found. Reading the second row in the Table we become even more aware of this patterning. A *SuperCluster* is a single instance of a set of *Clusters* typically juxtaposed to form a repeating pattern while *"Genre" or "SuperGenre"* is the term used to define repeated occurrences of the same type of *SuperCluster*. Why is this useful? The answer is that while there is a high probability of finding a *PhotoComplex SuperGenre* on a Migration Web page i.e. a semiotic unit made up of a set of juxtaposed photos relating to migrant workers, there is an almost near certainty that we will *never* find a *SearchEngine SuperGenre* made up of a set of juxtaposed *Search Engines* in *any* Website. In principle, this approach has applications in terms of predictability, for example, when attempting to identify the semiotic units associated with certain expressions of cultural diversity and their changing nature over time.

Appendix E: Glossary of the terms used in faceted approaches

Canon: A principle regulating practice in a given subject.

Idea: Product of thinking, reflecting, imagining, etc. got by the intellect by integrating with the aid of logic and/or what is directly apprehended by the intuition, and deposited in the memory.

Concept: A concept is the product of association of already deposited ideas. In other words it can be said as a perception about a thing by using many ideas. For instance, when a child sees a cow by using his imagination and intuition he develops many ideas like “it has four legs”, “it has one tail” etc. Later he can make a concept of cow in his mind by associating all these ideas. A concept is therefore a formation deposited in memory as a result of association with other percepts already deposited in the memory.

Isolate: A single component (‘ingredient’) of a compound subject [64]. Any idea or idea-complex fit to form a component of a subject, but not by itself fit to be deemed a subject. Isolates are distinctly different objects, which may be associated; some are merely distinguishable, though separable by various means such as quality, consistency, and operations. The relations among isolates are not static [76]. These are of two kinds: common isolate and special isolate (see below).

Common isolate: Isolates that can form components of several compound subjects going with each of all or almost all of the basic subjects. For instance, *Structure* is a component of many different subjects such as *Physical Structure*, *Chemical Structure*.

Special isolate: Form components of several compound subjects, which belong to a particular basic subject. For instance, *infectious* is a special isolate which can appear with *disease*.

Subject: The organized or systematized body of ideas within a field of interest or study. They are often organized in the form of academic disciplines.

Basic subject: Subjects which do not have isolate ideas as a component are basic subjects, for instance *Mathematics* [70].

Compound: A subject with a basic subject and one or more isolate ideas as components. In other words, it is a subject in which two or more constituent simple concepts are combined. For instance, *brain surgery*.

Complex Subject: A subject formed by coupling two or more subjects. The exposition of the first is in some way limited by the second. For instance, *Mathematics for physicist*.

Entity: Any existent: concrete or conceptual – that is a thing or an idea [70], which can have attributes, some of which can serve as a characteristic for division of the main class into groups. Any object or group of objects which has or may have existed, or which can be held to exist.

Characteristic: An attribute or any attribute-complex with reference to which the likeness or unlikeness of entities can be determined and at least two of them are unlike. In other words, it is the attribute

forming the basis of division in classification. *Language, Form and period are the common characteristics in the classification of literature.*

Train of Characteristics: The successive characteristics used to derive a chain of isolates.

Mutually exclusive: In the practice of mathematics, two events are mutually exclusive if it is not possible for both of them to occur at the same time. In the case of division of a universe, it is the characteristics, which must be mutually exclusive so that no two headings shall overlap. For example, if we choose to divide the places according to the countries name, we can choose a characteristic like "country name" and we can have the isolates like *India, Italy*, etc. but again we can't choose another characteristic, e.g. "political units", because it will also lead to the name of countries.

Group: Any sub-aggregate of entities formed by the division of the entities in a universe

Array: The set of mutually exclusive co-ordinate subclasses totally exhaustive of a class derived by its division according to any one characteristic at any one step in the progress towards its complete assortment and arranged in the preferred sequence.

Chain: A hierarchy of subclasses of decreasing extension and increasing intention derived by successive division.

Fundamental Category: A form or class of concepts, varying from subject to subject, into which isolates can be grouped, e.g.: matter or material, energy or action, organ property, space, time. Ranganathan postulates five fundamental categories: Personality Matter, Energy, Time and Space. Personality covers a particular occurrence (also called manifestation) of wholeness, for example, chemical compounds, plants and animals, and parts of them, languages, and religions: other fundamental categories are attributes of personality. The CRG (Classification Research Group, London) found a list of 13 helpful, substance (product), organ, constituent, structure, shape, property, object of action (patient, raw material), action, operation, process, agent, space, and time. Fundamental categories are useful as a provisional guide in approaching the analysis of a new field, providing an outline framework and giving guidance in suggesting possible characteristics which should not be overlooked [76].

Rounds and levels: Rounds and levels are tools to introduce or to allow more than one occurrence of fundamental categories. In Ranganathan's fundamental categories, any Energy facet can introduce a new round of Personality, Matter and energy facets. Personality, Matter, Space and Time can appear more than once in a round, these appearances are known as levels. For example, for the title **"***Ultraviolet Treatment of Bone Cancer in India in Summer 2007***"**, the facet sequence would be: *Medicine* [BS]; *Bone* [1P]; *Cancer* [1E]; *Ultraviolet Ray* [2P]; *Treatment* [2E]; *India* [S]; *2007* [T]; *Summer* [T2]. Where:

BS= Basic Subject
1P = First round of personality
2P= Second Round of personality
T2= Second level of time.
1E= First round of Energy
2E= Second round of Energy

As the disease *Cancer* is an action on personality *Bone*, it forms the first round of energy. *Treatment* is again an action on disease so it forms second round of energy. *Ultraviolet ray* is the personality which is being used for treatment so it deems to occur in second round. Similarly within a round every occurrence of personality, matter, space and time will be called as level. As in the above example, *Summer* is second level of Time.

Facet: The totality of the subclasses of a basic class corresponding to a single fundamental category and based on a coherent set of characteristics not leading to an organ or a constituent of the typical entity of the universe classified [64]. A generic term used to denote any component of a compound subject, also its ranked forms, terms and numbers [70]. Groups of terms derived by taking each term and defining it, per genus et differentiam, with respect for its parent class [76]. Or it can be said as a homogeneous group or category derived from the universe of entities or knowledge by applying a set of characteristics. It may also be seen as one part of a subject, situation, etc. that has many parts. In other words a facet is a hierarchy of concepts which are used to construct (single or compound) subjects in a given domain.

Focus: A generic term used to denote any isolate or any subject and also the number of any of them as well the name denoting any of them. Focus is a particular occurrence (also known as manifestation) of the basic class or of the fundamental categories concerned.

Facet Analysis: Analysis of a subject into its facets according to the postulates and principles stated for that purpose (Ranganathan). The essence of facet analysis is the sorting of terms in a given field of knowledge into homogeneous, mutually exclusive facets, each derived from the parent universe by a single characteristic of division [76].

Facet Synthesis: Synthesis of the focal numbers of a subject into class numbers according to the postulates and principles stated for that purpose. It is a process of creating new compound terms by combining facets.

Analytico-synthetic Classification: A classification which represents a subject by analyzing it into its fundamental constituents (isolates) and synthesizing class numbers for the subject out of the isolate numbers linked by appropriate connecting digits, e.g. Colon Classification

Enumerative classification: A classification made up by enumerating separately all the subjects to be included, e.g. the Library of Congress Classification.

Basic Class: A main class, i.e. a member of first order array of the universe of Knowledge, or a canonical or traditional subclass of a main class. A basic class is a unit member of the product classes obtained by dividing the universe of knowledge once.

Appendix F: Glossary of the terms used in Media Content Analysis

Media Content Analysis:
It is an empirical method for (I) systematic and inter-subjective understandable description of textual and formal characteristics (II) and for inquiring into social reality that consist of inferring features of a non-manifest context from features of a manifest written text and other meaningful matters [40][39][38].

Levels of analysis
The level of analysis indicates different units of the material, which are used for analysis. We distinguish between sampling unit, recording unit, content unit (coding unit) and context unit.

Sampling unit
The sampling unit consists of the selected material, which is used for the study (also called sample).

Recording unit
It consists of the selected *parts* from the sampling unit (i.e. one certain article from a selected webpage).

Unit of analysis
The unit of analysis indicates which elements from the sampling unit will be used for analysis. The unit of analysis can therefore be

- a whole newspaper,
- an article,
- photos within articles,
- statements within articles,
- frames within statements.

Content (or coding) unit
The content (or coding) unit is the interesting characteristic of the text on a certain level of analysis[29]. For example we ask for the speaker (opinion holder) on the level of statement.

Context units
„Context units are units of textual matter that set limits on the information to be considered in the description of recording units. […] the meaning of a word typically depends on its syntactical role within a sentence. To identify which meaning applies to a word from a list of dictionary entries, one must examine the sentence in which the word occurs. […] Here, the sentence is the context unit and the word is the recording unit. Sentences are the minimal context units for individual words, but sentences may not be enough. To identify the referent of a personal pronoun, for instance, an analyst may need to examine a few sentences preceding that pronoun. To judge whether a political commentary is positive or negative for a candidate, an analyst might need to examine even larger context units, such as a paragraph or a whole speech." (Krippendorf 2004: 101)

Codebook
The Codebook consists of the rules for coding: the categories, codes, definitions of codes.

Coding / to code
Coding / to code: The manual application of a codebook to texts is called coding. Humans read the content and select the appropriate codes from the codebook and fill out an excel-sheet or SPSS-file with these codes for every text.

[29] See example in the appendix of the report delivered in WP8 (deliver WP8.1).

Indicators
Indicators transform the theoretical assumptions into empirically investigable variables. Each indicator consists of one or more different variables.

Variables
The variables from the Codebook, which are further aggregated into indicators, are used for statistical purposes when responding to research questions. Each variable consist of two or more labels, which refer to certain codes.

Codes
Codes are numerical values. Those numerical values refer to certain information, which the texts are analysed for. The codes are necessary for further data analysis.

Claim / Statement
Claim is used in a more general sense of "statement" and is defined as the expression of a fact, a claim or an opinion by physical or verbal action in the public sphere.
A statement ideally consists of
1. Location of the statement in time and space (when and where has the statement been made?)
2. Speaker: the actor making the statement (who makes the statement?)
3. Form of the statement (e.g. political action, verbal statement etc.)
4. The addressee of the statement (at whom is the statement directed?)
5. The substantive issue of the statement (what is the statement about?)
6. Object actor: who is affected by the statement (For/against whom?)
7. The justification for the statement (Why should this action by undertaken?)

Frames
Frames are patterns of interpretation for different topics[30].

Research question
A research question is the starting point for every research, for theoretical studies as well as for empirical studies. Research questions are used both in natural sciences and social sciences. A research question sets out what the researcher is trying to get by his/her research. It defines clearly what the analysis should be about and indicates what the researcher wants to know. It also makes theoretical assumptions explicit.

[30] See Appendix A for an example.