AN END-TO-END APPROACH FOR DELIVERING DATA-DRIVEN STORIES

Sabino Metta and Alberto Messina

Rai Radiotelevisione Italiana, Italy

ABSTRACT

With the sheer scale of digital information now available, many journalists have recently started using data in order to tell compelling stories. Data journalism is becoming in fact a fundamental practice to increase information trustworthiness, obtaining full digital products exploitable on several platforms and improving user experience. Nevertheless, due to the complexity of the data ecosystem, this practice represents a challenge for any company involved in news production. Editorial staffs, in charge of extracting sense out of data to create newsworthy stories, need to be properly supported by targeted methodologies and efficient technological solutions. At this purpose, a prototypal workflow model is here presented. By retaining absolutely strategic the integration between proprietary and state-of-the-art tools, an experimental implementation of a toolbox and of an integrated platform is also described.

INTRODUCTION

In the modern Big Data age, information delivered by media is exponentially increasing and becoming impressively overwhelming. Leveraging social networks, search engines optimisation more than traditional networks, news bounce all around the world in a very few seconds. As a consequence, it is now commonplace to think that information is normally available and accessible in an easy and quick way to everyone. Nevertheless, *relevant* information (e.g. statistical data, relationships between persons playing a role in a story, etc.) is often delivered by newstellers in an implicit form. Most of the times, this additional information is delivered by a short text item and/or a video content (e.g., a chart) without any specific and stable structure or format, thus namely far from being *machine-readable, reusable and exploitable in the long term.* This means that, once the news has been delivered, such an information can be hardly retrieved and/or extracted for further investigations or insights.

At this purpose, in the last few years, the technologies for professional Big Data analytics have been representing a strategic necessity to make information resources available to professional journalists and media producers in a more effective and efficient way. The challenge lies in the ability of collecting, connecting, analysing and presenting heterogeneous content streams accessible through different sources, such as digital TV, the Internet, news agencies, social networks and media archives, and published through different media modalities, such as audio, speech, text and video, in an organic and semantics-driven way.

Since a few years, the sheer scale of digital information and technologies available has enabled new forms of production. In fact, many journalists have started using data in order to tell compelling stories. This relatively new practice, also called Data Driven Journalism (DDJ), is becoming a fundamental ingredient in order to increase information trustworthiness, obtaining full digital products exploitable on several platforms and finally improving user experience.

But together with the enthusiasm and the promising potentialities, this practice raises important management issues. From a media company perspective, especially for a public broadcaster, this innovative form of journalism can have a disruptive impact on the existing workflow, in terms of enhancement of the skills of its own employees and of the technological innovation needed. Obviously, a smart management should also take into account the business model continuously in evolution. Recently, a *Global Data Journalism survey*[1] has been launched in order to study into the current state of DDJ. Over 200 people responded, from 40 countries worldwide, including many from outside traditional newsrooms, including consultants, PR professionals entrepreneurs and academics. It should be noticed that broadcasters employed the smallest number (10%).

This paper reports about RAI's recently started experimental activity aimed at identifying an end-to-end workflow for delivering data driven stories. Such a workflow includes the design, implementation and integration of fundamental components related to the data refining, data analysis, structured and machine-readable story representation, open data management and advanced visualisation. This project defines and implements components and systems for professional information services that address these challenges with a uniform and holistic approach. At the core of the approach there is a set of artificial intelligence techniques and advanced statistical tools to automate tasks such as information extraction and multimedia content analysis targeted at discovering semantic links between resources, providing users with text, graphics and video news organized according to their individual interests. The system allows to define customized search profiles that are automatically and dynamically updated with the relevant contents found in the monitored information sources, which include Web feeds, television channels and specialized circuits such as the Eurovision News Exchange Network (EVN), or legacy archives.

RELATED WORKS

Recently, production facilities are more and more equipped with professional dashboards which implement complex artificial intelligence algorithms and technologies (AI). These dashboards are aimed at measuring and monitoring the behaviour of users while they are enjoying specific contents thus supporting editorial teams in their decision-making. So, in general, these technologies are intended to automate many of the time-consuming and labour-intensive tasks that are not directly related to the content, thus empowering journalists and programme authors in their work.

---

1 https://docs.google.com/forms/d/e/1FAIpQLSe_FVQdaMG6AU4oi_Z4LWHdcUE9p8fnHPDjpOVlk2sTVIuKZw/closedform

The usage of IT technologies in the newsroom is usually intended to automate and speed up specific processes within the news production chain. Recently, newsrooms started to employ technologies also for analysing generic 'data', namely documents, videos, tables, etc., with the purpose of creating new stories. In fact, data gathering, analysis and verification is supposed to improve the reliability and completeness of the delivered information. Clearly, data availability allows for creating visualisations and (possibly interactive) graphs thus eventually improving the user experience. These are the reasons for which all around the world this approach is getting ever-growing attention and interest even from academic parties (see the master degrees or the courses launched by several international universities, e.g. in Europe the *Máster en Periodismo de Investigación, Datos y Visualización* launched by the University Rey Juan Carlos of Madrid and the *ProfCert Data Journalism* launched by the University of Dublin UCD). In addition, a data-driven approach is a pivotal element within a workflow supporting the production of "full digital" contents. Nevertheless, it is absolutely common to observe an initial strong reticence showed by the editorial staffs in accepting and adopting the usage of technology for these aims. It is clear how it is not just a technology issue but, at the same time, also a cultural issue that need to be faced.

Despite this, in the last few years, the number of newsrooms which are delivering data-driven stories has been continuously increasing. Figure 1 contains a short selection of the main newsrooms.

In order to give an example of data journalism, we can mention *Immigration Reporter*[2] provided by NYT. An interactive map, see Figure 2, is aimed at showing how immigrant settlement has changed in each state over the time. The exact number of resident immigrant populations can be explored by the user. Immigrant populations are organized by colors assigned to each ethnic group. It can be noticed that the most substantial immigrant populations is near the Mexico border, south Florida, and the big cities.

The New York Times (NYT)[3] has been investing huge amount of resources in integrated technical solutions and aiming from the start to a full digital product. NYT has launched *The Upshot*[4], a new site featuring a combination of data-driven and explanatory reporting where analysis of the news and data visualisations are mutually combined. *BBC*[5] clearly sees the DDJ as a

Figure 1 – Some important examples of international editors involved in DDJ activities

[2] http://www.nytimes.com/interactive/2009/03/10/us/20090310-immigration-explorer.html

[3] https://www.nytimes.com/

[4] https://www.nytimes.com/section/upshot

[5] http://www.bbc.com/news

public service. They have focused mostly on visual journalism and on the creation of ad hoc apps. *The Guardian*[6] is absolutely one of the precursors (probably the first one) of this novel approach to journalism. They continuously work on improving the product quality. One of their most important killing factor is the establishment of perfectly integrated desk where multidisciplinary professional roles (not only journalists but also developers, designers, etc.) collaborate since the start on each new story creation. *ProPublica*[7] stands out for the high technical

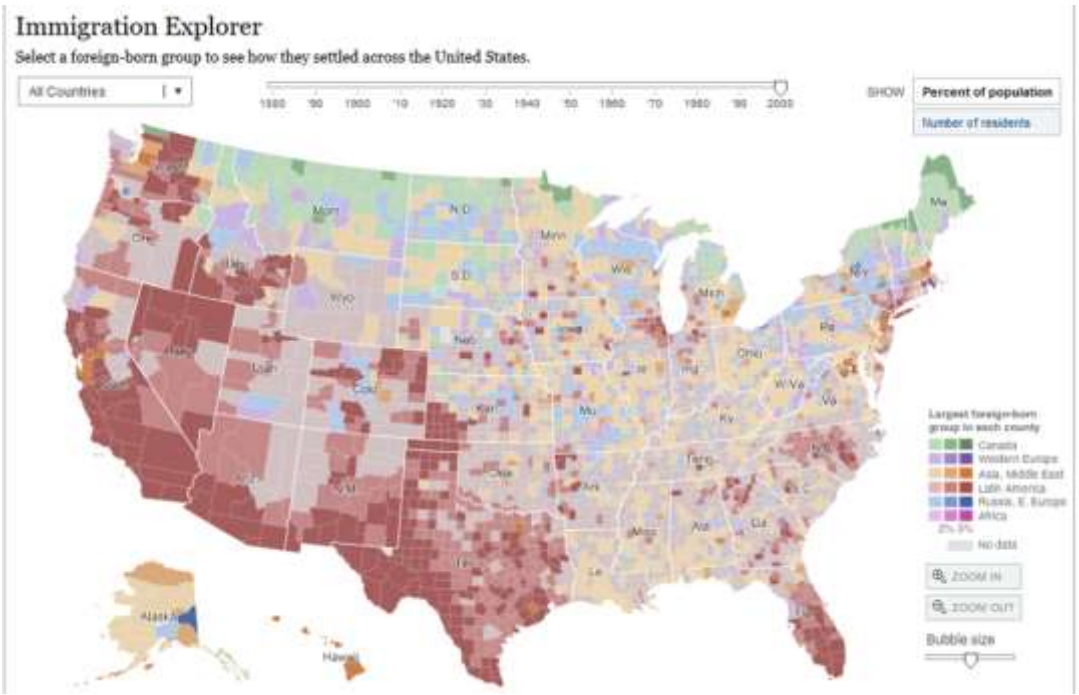

Figure 2 – *Immigration Explorer* by NYT. An example of storytelling with data.

competence of its journalists which normally use data as source for the narration. *La Nacion*[8] also experiments Open Source and Open Data Journalism. They have been relying on crowdsourcing for supporting and spreading their projects. At *ICIJ*[9] investigative journalism is produced though data. An important characteristic is represented by their international collaborations (see the “Panama Papers” investigation which recently won the Pulitzer prize) with a rich network of reporters in order to exchange relevant data. In investigative journalism, the workflow and the toolbox are strongly affected by the “threat-model”. Journalists have to select those particular tools they can use by carefully evaluating the dangers and the threats implicated in the specific investigation. Eventually,

[6] https://www.theguardian.com/international

[7] https://www.propublica.org/

[8] http://www.lanacion.com.ar/

[9] https://www.icij.org/

The Texas Tribune[10] started producing very local works thus delivering ad hoc news to citizens. Now they are trying to expand to international perspectives. They have been experienced new forms of storytelling by putting data as pivotal element.

These few examples have been gathered and studied to identify the different approaches and objectives among the most important newsrooms. In fact, despite DDJ is a worldwide practice, newsrooms normally don't share a standard workflow neither a well-defined toolbox nor an integrated platform. At this purpose, since the last few months RAI started a methodological and technological project for including DDJ activities in its internal production chain [1,2]. Major details will be provided in the next section.

END-TO-END APPROACH

In this section, a prototypal workflow model for supporting DDJ activities is presented. It is also described the experimental implementation of an integrated platform where proprietary, state-of-the-art and in-house tools are integrated. The project is supposed to improve the quality of the final editorial product and also to optimise the current organizational workflow.

In a DDJ work, we observe how journalists usually do a hectically recursive sequence of actions thus following often a nonlinear and sometimes very complex workflow. Distilling out the core set of actions implied in the delivering of a DDJ work, we can instead identify a 'linear' process flowing from "data" to "publication".

data → analysis → story → publication

In general, 'data' do not refer uniquely to numbers, but mainly to structured and machine-readable information. In the process of finding the story journalists often need to gather as much as possible information, to share documents, etc. They can rely on several existing tools, e.g. *Google Doc*, *Dropbox Paper*[11], *Workflowy*[12], etc. Nevertheless, from a company perspective it is clearly a strategic practice to rely on internal solutions thus avoiding any information leakage issues. By analysing some examples of international best practices we can outline, at least, two main aspects publishers have to deal with. First, it is absolutely necessary to have tools (mostly already available in commercial, open and/or free option) for managing data in terms of retrieval, refining, modelling, analysing and visualising. Then, the newsroom strongly needs an interdisciplinary team where different skills and expertise (not only in journalism but also in programming, math, statistics etc.) are mutually collaborating for delivering the story. One of the key factors is the strong engagement of all the components of the group since the first draft of a story, and not in a later and secondary stage once the story has been already written.

In order to keep the editorial flow as much broad and flexible as possible, at RAI we focused on some fundamental potentialities of data: they can be themselves the origin of news, of a story, or of generic informative content; they can represent an accurate and

[10] https://www.texastribune.org/

[11] https://www.dropbox.com/paper

[12] https://workflowy.com/

verifiable description of possible semantic aspects inherent a story; they are the fundamental objects of visualisation. According to these conditions, the project is aimed at facing some important aspects of the complex DDJ ecosystem in a stepwise manner. At the current stage the project primarily intends to address *search & retrieval* and *data analysis* problems. Later on, the project will approach the problem of curating and managing the journalistic data-base, i.e. where all relevant metadata associated to analysed data are indexed and shared. Eventually, the project will integrate the overall newsroom workflow in a unique platform. Such a methodological and technological support is aimed at valorising every single story. Leveraging this unique platform, journalists and editorial teams will be able to collaborate together on the story editing adapting it to the target audience they want to reach.

The implied technological aspects involved in the project are summarised in the following list:

- **Semantic Web Technologies,** to support automatic metadata extraction from textual sources. Specific and/or generic *ontologies* (namely the formalizations of concepts and relationships used to describe and represent an area of concern) will be adopted to represent the relevant domains of the journalistic investigations;
- Available **Open Data** (together with Linked Open Data) will be integrated and made available to support the data-driven journalistic work;
- **Data Visualization & Infographics** for the automated production of interactive charts, e.g., visualization of time-series data and/or timed content in the database.
- **(Multiplatform) Content Management System** integration with existing internal CMS for the final publication.

It has to be noticed that the technologies developed and integrated within this project can be reused internally also for other tasks than the journalistic ones.

Figure 3 contains the main methodological and technological components RAI has currently focused on to produce data driven stories.

By using a concept mapping tool (e.g. **Freemind**), the editorial team can conceptualize and model relevant information referred to a specific story. Leveraging the same integrated platform, the team can access, store and distribute relevant information as open data (**CKAN**). According to specific needs, the team can also rely on an in-

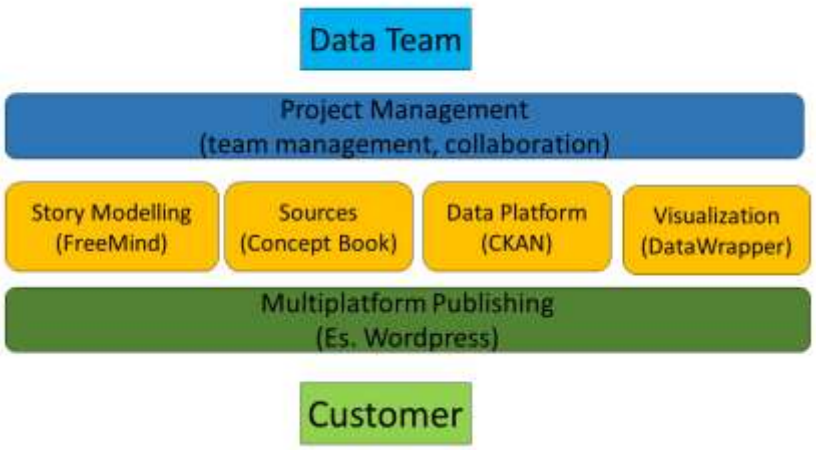

Figure 3 – Architectural overview of RAI project

house tool (**Concept Book**) for automating tasks such as information extraction and multimedia content analysis. Then, by using a data visualization platform (e.g., **Datawrapper** *Google* tools/*infogr.am*[13], etc.) simple charts can be created and embedded in a content management system (e.g., **WordPress**). In this way, the data-driven production chain is supported from the data harvesting up to the publishing stage.

The **Data Team** is a group of people with multidisciplinary skills, journalistic and/or technical. Technical members support journalists in finding/harvesting/processing data and participate to the story creation/curation. On the other hand, journalists set the editorial line, develop the story, assess data relevance. The result is that the synergy and the delegation increase the efficiency In addition, cross-collaboration gives birth to unexpected explorations.

RAI **Concept Book** is a portal for professional information services based on artificial intelligence and advanced statistical tools [3,4,5,6]. These are used to automate tasks such as information extraction and multimedia content analysis (Figure 5). The system allows to define customized search profiles that are automatically and dynamically updated with the relevant contents found in the monitored information sources. Its flexible architecture allows for the addition of further sources (Figure 4).

Comprehensive Knowledge Archive Network, also referred as **CKAN**, is an Open Source platform used by many organizations as a platform for open data publication. RAI uses it as a platform for data journalism production. In fact CKAN allows easy integration with other CKAN-based open data repositories and provides updates of data automatically. Then it is extensible via plugins.

Integrating the aforementioned tools within a unified workflow is absolutely not trivial due to the complex ecosystem of proprietary and commercial software solutions currently adopted by the internal staff. As an example of such integration. the data source sets selected from the RAI Concept Book become plain CKAN datasets. On the other hand, datasets harvested by CKAN are classified and analyzed by RAI Concept Book toolbox.

[13] https://infogr.am/

Finally, in order to illustrate the potential of the methodology and the main functionalities of the tools implemented, some sample stories have been created and curated. These stories have been published on the WordPress platform (WP). By installing specific WP plugins it has been possible to include in the story board also charts, tables and timelines previously produced in the described experimental production chain.

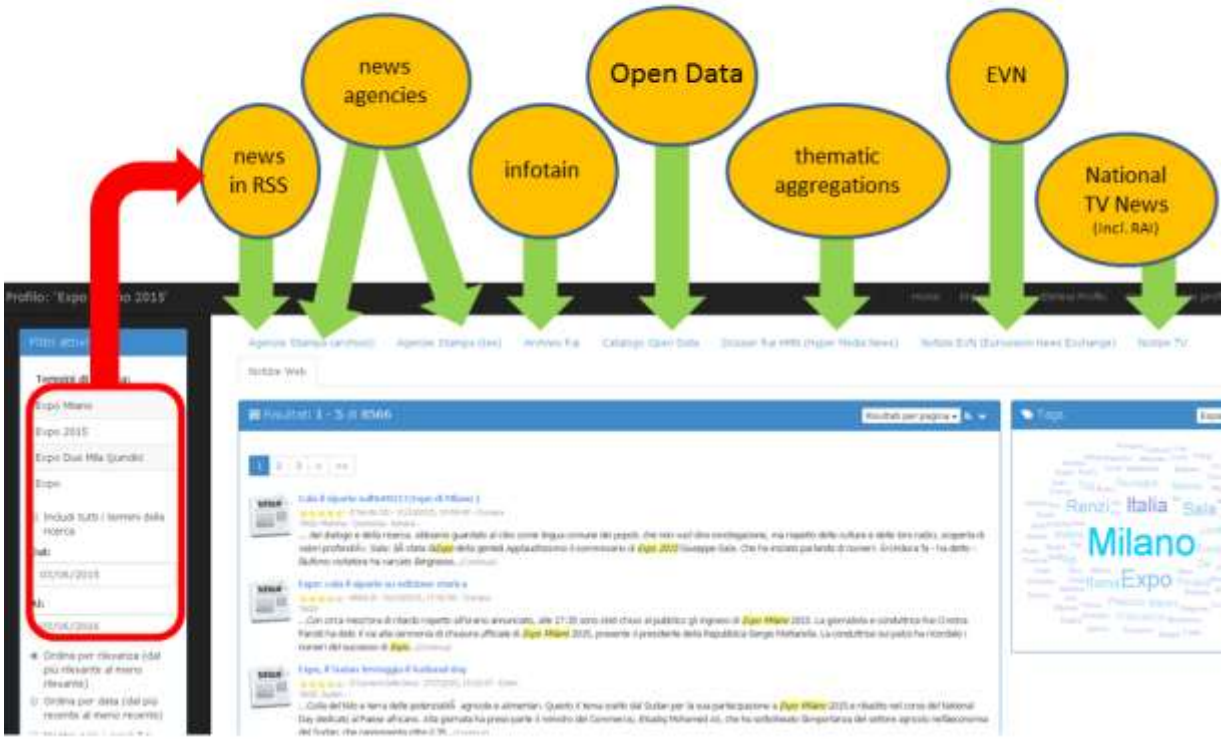

Figure 4 – Example of multiple sources integrated in the system

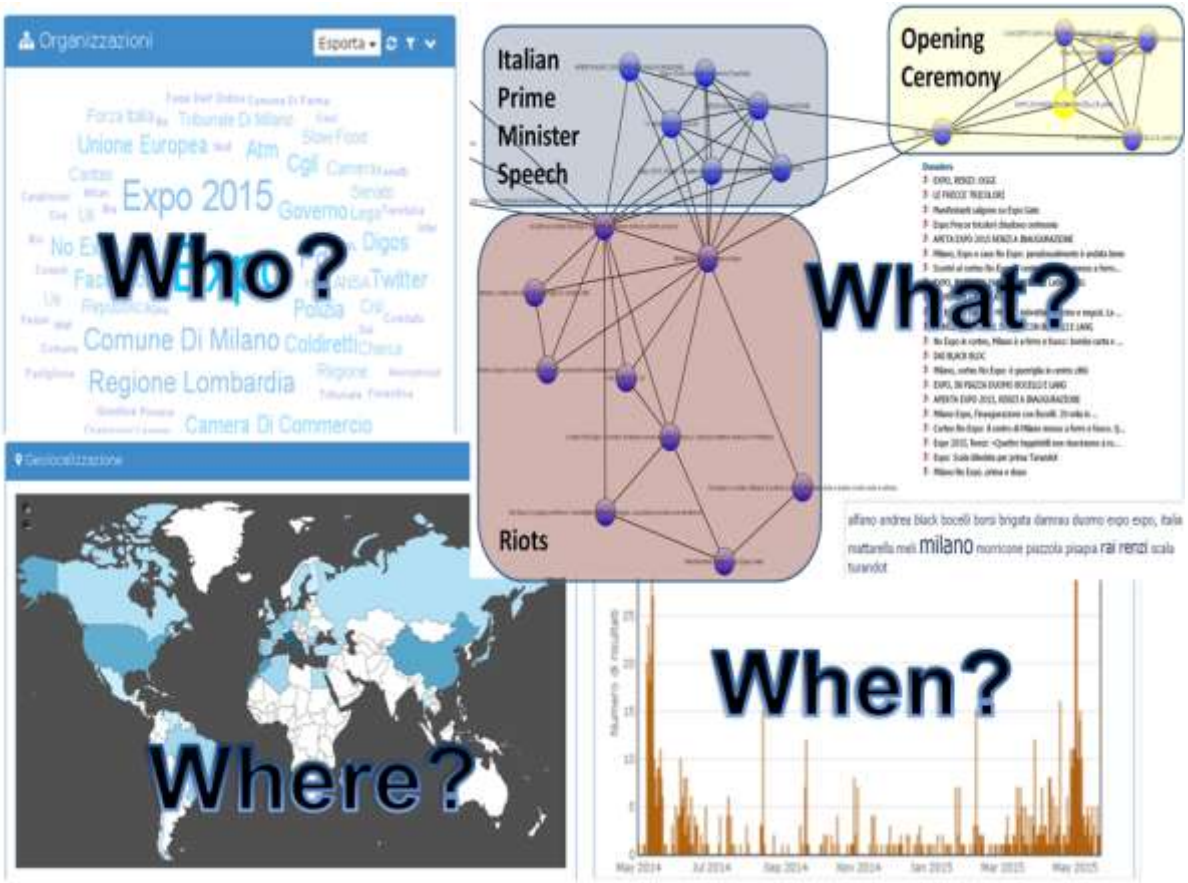

Figure 5 – Example of analysis results performed by Rai Concept Book

CONCLUSIONS

In the last few years Data Driven Journalism (DDJ) has become a fundamental practice to increase information trustworthiness, obtaining full digital products exploitable on several platforms and improving user experience. Nevertheless, due to the complexity of the data ecosystem, this practice represents a challenge for any company involved in news production. At this purpose, RAI started an experimental project aimed at facing some important methodological and technological aspects involved in DDJ activities. A prototypal workflow model has been designed and its implementation started. The work focused on the implementation of a toolbox and of an experimental platform where in-house software and SotA tools have been integrated. In addition to the technological aspects, the project clearly identified the need of a multidisciplinary team where collaboration is supposed to start since the beginning of the story creation.

RAI is now planning to focus its research efforts on a tighter integration with Semantic Data and Open Data towards more semantic-oriented structuring. Some experimental activities are aimed at exploring the possible strategic role of Visual Search technologies in DDJ. Eventually, in order to increase the user engagement, RAI is going to study enhanced professional dashboard for publishing data driven stories on second screen devices.

REFERENCES

[1] Giulia Dezi, Giorgio Dimino, Maurizio Mazzoneschi, Alberto Messina, Sabino Metta, Giuseppe Mondelli, Maurizio Montagnuolo, EBU MDN Workshop, 2016

[2] Sabino Metta, Open Data e Data Journalism, Conferenza TAL & Open Data, 2014

[3] Alberto Messina, Maurizio Montagnuolo, Riccardo Di Massa, Roberto Borgotallo: Hyper Media News: a fully automated platform for large scale analysis, production and distribution of multimodal news content. Multimedia Tools Appl. 63(2): 427-460 (2013)

[4] Maurizio Montagnuolo, Alberto Messina: The RaiNewsbook: browsing worldwide multimodal news stories by facts, entities and dates. WWW (Companion Volume) 2012: 389-392

[5] Alberto Messina, Maurizio Montagnuolo: A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval. WWW 2009: 321-330

[6] Alberto Messina, Roberto Borgotallo, Giorgio Dimino, Daniele Airola Gnota, Laurent Boch: ANTS: A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis. WIAMIS 2008: 219-222