

# Tools and resources supporting the cultural tourism

Sassolini Eva, Cinini Alessandra, Stefano Sbrulli, Picchi Eugenio

Organization: Istituto di Linguistica Computazionale “Antonio Zampoli”  
Address: Via Moruzzi, 1, Area della Ricerca di Pisa  
Postal Code/City/Country: 56124, Pisa, Italia  
Topics: Repurposing Grey literature  
Adapting New Technologies  
Telephone: 050 315-2759/2859  
Fax: 0503152839  
Email: {eva.sassolini|alessandra.cinini|stefano.sbrulli|picchi}@ilc.cnr.it

## Introduction

The diffusion of internet and the information technologies are creating continuous information flows. There is a widespread awareness of the added value and of the role that the Web has in dissemination, exploitation and promotion of the cultural tourism, especially in a country like Italy, where the cultural heritage is very important. Moreover, an open philosophy causes problems of authoritativeness in the production of contents because it is characterized by a strong interaction among users thus creating a distance between knowledge and communication. The spread of internet has brought the significant changes of communication paradigm. Nowadays the competition decreases among contents, even among from sources published in potential competition with them. In network logic, all nodes are interdependent and represent a single large hypertext. The proliferation of paths boosts a free circulation of ideas and can bring out most interesting contents[1].

## The Projects

Methodologies, strategies, resources and tools created in our research group allowed us to take part in some national initiatives:

- “On-line dissemination of the historical artistic and landscaped, regional heritage” (*WeBasCH*): The project was born within the framework of collaboration between the ILC-CNR Pisa and the APT Basilicata (i.e. Agenzia di Promozione Territoriale della regione Basilicata) to experiment and implement strategies for promotion and dissemination of regional heritage.
- “*SmartCity*: new solutions for content engineering and ambient intelligence as support of cultural tourism” is a Tuscany region project ("POR Creo" session), funded by the European Community (FESR). The specific project purpose is to allow important steps forward in terms of productivity, versatility and adaptability of digital content, both textual and multimedia, through knowledge engineering techniques. Particularly we have developed and tested tools for a better preservation and enhancement of cultural heritage, identifying methodologies and solutions to meet new demand of cultural spaces in particular for tourist purposes. Partners of this project were ILC-CNR and a consortium of companies: Space, Rigel Engineering and Meta.

The specific goal of both projects is experimenting and implementing strategies for promotion and dissemination of regional heritage through an effective communication style. By exploiting potentialities of the Web 2.0, we offered a set of chosen documents and cultural routes that increase knowledge of historic and landscaped resources. The aim is to join maintenance and preservation activities and enhancement and promotion of the cultural heritage initiatives with an increasing

opening to the general user[2].

## **Text material for cultural tourism**

The cultural tourism in Italy aims at identifying town of artistic interest or cultural tours in wider areas linked by historical events or traditions as the "Via Francigena", " Parco Archeologico Storico Naturale delle Chiese Rupestri ", "the wine roads", etc.. In this context, information available is in continuous evolution and so various that the database should be continually updated. The updating process of textual resources is an open question. The manually generation of resources is expensive and it requires the review of human experts. The manual approach is prone to errors of omission. Moreover, an approach based on an automatic updating of resources is not efficient because these are constantly evolving, this reason advises against an exclusive use of automatic applications. Especially when it comes to tourism, for which the Web offers constantly new ideas: a new park, an archaeological site discovered, etc..

There are many European initiatives within Cultural Heritage that aim to develop knowledge and enhancement of digital cultural heritage that have been undertaken. Among these, "Minerva" and "Michael" have been coordinated by the Italian Ministry for Cultural Heritage. Minerva has developed a platform of guidelines and recommendations, which are shared by European member states, for the digitization of cultural heritage and its network access. Similar information can be considered the reliable source of knowledge, especially because they were created with specific objectives:

- ✓ accessibility and visibility improvement of European digital cultural resources;
- ✓ support development of the European Digital Library for a better access to the cultural resources;
- ✓ contribution to increasing the interoperability among existing networks;
- ✓ promoting the use of digital cultural resources by business and citizens.

However the creation of this kind of resources is also referred to written text material such as brochures and web advertisement, that often is difficult to find, and retrieve. Such data regard an important source of information but since it tends to be original, recent and ephemeral can be considered a typical material of Grey Literature.

## **Specific strategies for texts acquisition**

We have dealt the creation of the text corpus with different strategies in the two projects. Different needs of each project have guided the choices of the best acquisition strategy and all text material has collected in two steps:

1. We built a starting text corpus (hereafter C0) which was exploited to generate new linguistic resources and enriching the ones available from TextPower<sup>1</sup>[3] project.
2. On the basis of these resources, we enlarged C0 to create the reference corpus.

## **CO in *SmartCity***

The first acquisition strategy in *SmartCity* project concerned the creation of a domain corpus "Empoli e dintorni", composed of historical, artistic and tourist documents related to the specific geographical area. Among activities of the project it has been possible collaborate with the library of Empoli and Cerreto Guidi for the retrieval of textual material.

On a first group of documents provided by the library of Empoli, we attempted to select material of interest in order to deal all the most important themes of Empoli and surroundings. On the basis of

---

<sup>1</sup>Project made for building of terminological resources and enrichment and annotation of textual material

these texts we generated the Training Corpus whose size is almost 110,00 occurrences of words.

## CO in *WeBasCH*

A different approach has been followed about *WeBasCH* project where all text material has retrieved on-line. The APT Basilicata provided us with a list of institutional websites. The items of the list more related to historical, artistic and landscaped regional heritage were selected and browsed by using automatic spiders and parsers, for the creation of a text corpus. Much of documentation available in internet can be assimilated to the “grey literature”, since it has been produced by the authors and institutions outside publishing, particularly the websites of regional entities.

## Text corpus to linguistic resources

C0 was exploited to generate new linguistic resources and integrating the ones already available. All textual material obtained has indexed and, after a tagging phase with the PiTagger<sup>2</sup> tool, we could identify all lemmas and relative POS in each document[4].

PiTagger associates each word to the related lemma by using the morphological component of the Italian language PiMorfo<sup>3</sup>. Then it solves the ambiguities by following a statistical approach and with the help of a training corpus.

Later on, all *multiword* expression (MWE) were extracted from C0 by exploiting pattern matching techniques. Typically for the Italian syntactic construction, the most productive linguistic patterns are N-preposition-N and ADj-N/N-ADj. Statistical algorithms analyze the distributions frequency of each pattern identified. On the basis of results we extracted a set of semantically relevant terms and concepts for the cultural heritage domain. The analysis of the collected texts by means of linguistic tools (morphological engine and tagger) is fundamental for productive application of the statistical functions of extraction[5].

We exploited enriched text material, that composes C0, to build weighed domain lexicons besides. Starting from a small set of relevant pivot terms a lexicon is obtained by means of *mutual information* criteria. Statistical algorithms analyze and weigh the frequency of the co-occurrence of each word with the pivot terms. The domain lexicons can be used to evaluate the relevance of a document for that domain, in this case it is most important to establish a minimal threshold.

## Reference (text) Corpus

Since textual documentation collected didn't cover exhaustively any events and cultural resources, we tried to increase the amount of textual material available, in order to find a maximum coverage of information and knowledge. In both projects we used a new search strategy of text material retrieval, namely, we used extracted knowledge and specialized crawlers that work on a bulk of text material available on-line. In a next phase we have developed other tools that, by using specific semantic filters, make a ranking the documents and evaluate their relevance with the specific domain. All extracted documents were joined in C0 corpus to build the reference (text) corpus.

The (text) reference Corpus created in SMARTCITY project consists of 2219 text units:

- 1634 units come from the material provided by the library of Empoli and Cerreto Guidi, about 1.000.000 (997299) words;
- 585 units are related to texts retrieved on the Web, little more than 850.00 words (857355).

---

<sup>2</sup> PiTagger is an important component for text lemmatization and tagging and constitutes a software module of PiSystem: integrated system for processing of textual and lexical materials.

<sup>3</sup> PiMorfo: system for morphological analysis of the Italian language.

At the end of the project the specific Corpus has been reorganized in 650 documents in XML format.

In the *WeBasCH* project, the (text) reference Corpus instead is constituted of almost 2 million and half words. As it described, the creation required two phases, in which we retrieved documents from the Web and built the reference (text) corpus.

## DBT-Faccette

The intelligent browsing system of text, named "DBT Faccette", is a customization of the categorization system used in librarianship. Its uniqueness lies in possibility of exploiting the semantic relevant elements ( or "micro semantics") identified in the text for suggesting a further search. This feature makes most important the availability of an annotated corpus.

The text corpus is originally a set of text files. Inside these texts specific tools have identified annotations of various kinds: information related to words, phrases and piece of text. However it is necessary to organize the annotated texts in a coherent way and orderly, for a better storage, management and expansibility over time.

The corpus becomes a collection of digital sources containing accurate semantic annotations and necessary information to managing of the textual sources.

A useful way to imagine a good browsing system is to assume the final usage scenarios. Many developers often tend to adopt "reference systems" implicit, made of the individual experiences. Since the "reference system" of a user is not necessarily identical to ours, it's easy to fall into misunderstandings harmful to the design of a complex product.

The browsing system is able to exploit knowledge extracted from textual material in several ways:

- expanding the search by exploiting the search for MWE and then offering suggestions to research incomplete or vague. For example, by proposing the query "Pontormo" are returned as results also the occurrences of "Jacopo da Pontormo", "Jacopo Carucci" and "Pontormo" (Fig. 1);

The screenshot displays the 'Ricerca con DBT Faccette' interface. At the top, a search bar contains the query 'Pontormo' with a 'Cerca' button. Below the search bar, the results are organized into a table with columns: 'Ricerca', 'Pontormo', 'Faccette', and 'Chiudi'. The 'Ricerca' column shows the search term 'Pontormo' and the document title 'Jacopo da Pontormo, diario: introduzione'. The 'Pontormo' column shows the search results, including a list of documents and a detailed snippet of text. The 'Faccette' column shows a list of related terms, including 'Jacopo Carucci' and 'Pontormo'. The 'Chiudi' column shows a 'Trovati: 313' count. The snippet of text in the 'Pontormo' column is highlighted in red and contains the following text: 'TIT: Diario di Jacopo da Pontormo - TIT: E sto così senza sapere quello che a essere di me. (Diario, 1 aprile 1555) Un mito. O, almeno, la fantasiosa elaborazione di una [...] filigrana [...] Pontormo - personaggio e simbolo [...] solo per [...] [...] arricchendo un altro mito quasi negli stessi anni del Pontormo - tutto ciò è drammaticamente concentrato sull'ossessione [...] il manoscritto è anepigrafo, autografo [...] note non [...] che sembrava una sentenza (Il lunatico Pontormo), a rinforzare di amentità l'aura mitica [...] tradizione vasariana, nella lettura di Cecchi il Pontormo diveniva una specie di prefucinato matto delle giuncale [...] sto rinchiuso come la midolla). In Pontormo, semmai, veniva a sovrapporsi - e [...] nel grottesco. Aprendo il diario del Pontormo, quindi, bisogna aver presenti alcuni punti [...] della banale registrazione dell'ipocondria. In Pontormo (come del resto nella sua pittura) accade [...] un colore necessario. Il testo che Pontormo redasse fra il 1554 e il 1556 deve [...] linea dell'orizzonte. Nonostante questo, Pontormo ha evidentemente fiducia nella scrittura: che è [...]

Figure 1: some results for query "Pontormo"

- improving the correlation process among documents identified, by using specific linguistic resources;
- improving the ranking of results in case of answers classification of responses;
- better organizing the display of the results. The system shows in reply at the queries a variety of contexts. In these results are highlighted all relevant entities or "micro-

semantics". In this way, the next search refinement is facilitated.(Fig. 2);

Ricerca :

stile barocco/centro storico/fmadonna del carmine

Ricerca

Faccette

Chiedi

decescente - alfabetico

Trovati: 33

basilicata 9	stile barocco 9	madonna del carmine 9	centro storico 9	matera 8
salvatore sebasto 8	scultura 7	iusco 6	anna grelle 6	pietro 6
de luca 6	presbiterio 6	chiesa 6	madonna del rosario 5	giacomo racioppi 5
potenza 5	lucania 5	ermanno loesch 5	madonna con bambino 4	gentilizio 4
padova 4	tela 4	angiol 4	annunziata 4	fonte battesimale 4
pallotto 4	altare maggiore 4	sansverino 4	santa maria 4	primo piano 4
sant'antonio 4	pittura 4	capella laterale 3	pietrafesa 4	legno policromo 3
padula 3	greci 3	napoli 3	puglia 3	angelo lucano 3
lucia 3	rosario palese 3	larotonda 3	giovanni di gregorio 3	giovanni battista 3
anime purganti 3	crocifisso ligneo 3	benevento 3	chiesa madre 3	capella 3
sculture 3	san giuseppe 2	misteri del rosario 2	san giovanni battista 2	san lorenzo 2
garibaldi 2	salerno 2	donato 2	leonardo 2	genzano 2
croce latina 2	san rocco 2	trattati 2	torre quadrata 2	palazzo ducale 2
anfitreato 2	stigliano 2	sala consiliare 2	del tufo 2	convento francescano 2
tricarico 2	arti grafiche mota 2	navata unica 2	complesso monumentale 2	altare centrale 2

basilicata.cc\_82

1

Un'altra piaga fu l'epidemia della "spagnola", malattia infettiva, che provocò numerose vittime. Nel **centro storico** si può ammirare il Castello, situato nel punto più alto del paese: si tratta più che di un castello di un palazzo a pianta quasi triangolare, con

2

volta a crociera. Interessante è la chiesa madre dedicata a Santa Maria Assunta, costruita con le pie elargizioni dei fedeli, in gran parte proprietari terrieri. Ha una facciata in **stile barocco** con campanile rifatto nel 1866 in seguito al crollo avvenuto durante il terremoto del 1857. Nell'interno a tre navate con altare maggiore in marmi policromi sono conservati un coro ligneo

3

marmi policromi sono conservati un coro ligneo del 1793 opera di falegnami lagonesesi, alcuni dipinti tra cui quelli della Madonna del Rosario del 1788 e dell'Assunta, le statue settecentesche della Madonna del Carmine, di San Giuseppe e Sant'Antonio. Nelle vicinanze della chiesa madre è situata anche la capella dell'Annunziata, che chiusa per alcuni anni è stata riaperta recentemente. La Madonna

4

da: A.P.T. Basilicata da Visitare: **Centro Storico** ( la Porticella ) / Chiesa Madre / Capella della Madonna della Sella / Palazzo Settembrini / Vasche di Sant'Alessio / Villa Imperiale / Lungomare Manifestazioni ed eventi: 19 marzo Festività

5

/ Lungomare Manifestazioni ed eventi: 19 marzo Festività in onore di San Giuseppe (Patrono) 13 giugno Festività in onore di Sant'Antonio 16 luglio F estività in onore della **Madonna del Carmine** Altri Eventi (segnalati) . agg. al 25/05/2008

6

trascoro nell'abbazia alcuni anni della sua vita), feudatari di Puglia e Basilicata le concedevano terre, casali e chiese. Il patrimonio era vastissimo ed ancora oggi il **centro storico** coincide in buona parte con il sito dell'insediamento monastico. Intorno al monastero sorse il casale per la residenza dei coloni a servizio della comunità. La fine del

7

eretto dai francescani nel 1688 si trova (fig. 9) un lavabo sostenuto da un capitello, proveniente dal chiostro medioevale. Dalla sacrestia si passa nella capella della **Madonna del Carmine** in cui si trovano dei bassamenti a stampella del XVIII secolo, provenienti dalla chiesa benedettina, un'altara d'altare di legno dorato e intagliato del XVIII secolo, una

8

l'ingresso del Palazzetto del Vicario con un interessante stemma cardinalizio. L'interno della chiesa (fig. 11), ad una navata con capelle laterali, fu completamente trasformato in **stile barocco** nel XVII secolo dai francescani. Entrando, a destra, sull'altare della capella di S. Vito c'è una bella scultura lignea del santo (sec. XVII).

vacanzeinbasilicata.it\_7

Figure 2: example of queries refinement

An example of browsing in WeBasCH shows that some material constitutes examples of gray literature. For example, the search of "premio" proposes the site of Aliano, particularly the web-page where planned activities for the year 2008 are displayed. This material would be lost over time, even though it provides important information about the natural and cultural heritage of a region.

[illegible]

**Figure 3: document retrieved in WebBasCH**

In our approach, the creation of linguistic resources is designed to the development of navigation and information retrieval systems, that are able to exploit them. These tools capture, organize, classify and distribute the information in according to the desired objectives. In an open domain, as the Cultural Heritage, information can rarely be classified just with hierarchical criteria. An approach based on principles of "semantic similarity" is more efficient. This approach allows to link information crosswise, seemingly belonging to different categories, but that match to the same

informative need.

The experience in the treatment of large amount of data has not only allowed the refinement of extraction tools for semantically relevant information, but also the creation of terminological resources toolkit.

The more a text is "enriched" with annotations, the better it can be processed by tools for analysis, categorization, browsing and Information Retrieval. The system is not limited to the identification and classification of entities; it also identifies the particular relations between the entities involved.

## Conclusion

Our work proposes to overcome the traditional categorization systems and their rigidity, by means of a set of open and adaptive terms classes, which can guide the end user in refining of his/her search.

Such knowledge systems are valuable support on the one hand to networks of e-participation and e-government, on the other hand they offer more information and better performance.

## References

1. Spadoni F., Tariffi F., Sassolini E., (2011). SMARTCITY: Innovative Technologies for customized and dynamic multimedia content production for Tourism applications. In: EVA 2011 Florence Electronic Imaging and the Visual Arts. (Firenze, 4-5-6 may 2011). Proceedings, Cappellini Vito (ed.). Pitagora Editrice Bologna, 130 - 135.
2. Granieri, G., Et Al., (2009) Linguaggi digitali per il turismo. Edizioni Apogeo, November 2006.
3. Picchi, E. Et Al. (2009). "Text Power": tools for Cultural Heritage. In Proceedings of in 4th International Congress on "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin", IMC (CNR) Rome, Italy, pp. 277--278.
4. Picchi, E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian. In Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex '94, Amsterdam, The Netherlands.
5. Picchi E., Et Al., (2004). Linguistic Miner. An Italian Linguistic Knowledge System. In: LREC Fourth International Conference on Language Resources and Evaluation (Lisboa-Portugal, 26-27-28 May 2004). Proceedings, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silvia (eds.), 1811 – 1814.