A Digital Archive of Research Papers in Computer Science

Manuela Sassi*, Gabriella Pardelli*, Stefania Biagioni, Carlo Carlesi**, Sara Goggi***

* Istituto di Linguistica Computazionale "Antonio Zampolli"

** Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"

manuela.sassi@ilc.cnr.it, gabriella.pardelli@ilc.cnr.it, stefania.biagioni@isti.cnr.it, carlo.carlesi@isti.cnr.it, sara.goggi@ilc.cnr.it,

Abstract

This paper presents the results of a terminological work conducted by the authors on a Digital Archives Net of the Italian National Research Council (CNR) in the field of Computer Science. In particular, the research tends to analyse the use of certain terms in Computer Science in order to verify their change over the time with the aim of retrieving from the net the very essence of documentation. Its main source is a reference corpus made up of 13,500 documents which collects the scientific productions of three CNR research Institutes. They are ISTI (Institute of Information Science and Technologies), IIT (Institute of Informatics and Telematics) and ILC (Institute of Computational Linguistics), all of them born from the "Centro Studi sulle Calcolatrici Elettroniche (CSCE)" and now belonging to the CNR Department of Information & Communication Technologies and Cultural Identity.

This study is divided in three sections:

1) an introductory one dedicated to the data extracted from the scientific documentation: the data have in common the use of some terms proper of the Computer Science lexicon although these term belong to different branches (Linguistics, Informatics and Telematics);

2) the second section is devoted to the description of the contents managed by the PUMA (Publication Management System) system;

3) the third part contains a statistical representation of terms extracted from archive: some comparison tables between the occurrences of the most used terms in the scientific documentation produced by the three Institutes will be created and diagrams with percentages about the most frequently used terms will be displayed too. Lastly, indexes and concordances will allow to reflect on the use of certain terms in this field and give possible keys for having access to the extraction of knowledge in the digital era.

1. Introduction

Nowadays the need of retrieving the great amount of digital knowledge available on the web is ever more important: the vast majority of this knowledge is conveyed by means of textual material stored in scientific documentary repositories and digital archives. This documentary word preserve inside the wealth of far-off terms belonging to the past which have often fallen out of use as well as a more recent terminological production derived from the feverish need of coining new terms, a very common trend in certain branches.

"Documentation is a relatively new field that deals with the collection, analysis, assignment, classification and storage of documentation in order to make the information retrievable for various uses, users, and purposes. Like terminology, documentation is cross-disciplinary in the sense that it can be applied to any branch of science or human activity. It is practice-oriented because its aim is to provide information users with suitable documents. Documentation centres also disseminate information in the form of secondary (bibliographies and dictionaries) or tertiary (bibliographies of bibliographies) publications" (Cabré, 1999: 1).

The technical-scientific vocabulary represents a significant part of the lexicon of any language and plays an important role in the information exchange between the experts of the field and the users of the Internet community, the new market of knowledge where people producing informative contents can meet – by means of a common lexicon - those who are actually looking for that specific information. Information dissemination on the World Wide Web contributes to make terminology ever more important: in particular, the use of certain specific terms which become the keys for retrieving information. Temmerman says: "Terms are the engine of any comprehension process since they point out a new comprehension as well as an old comprehension, and for the classic terminology terms denote concepts. (Temmerman, 2000: 228).

The analysis of scientific documentation in the field of Computer Science is especially relevant because its continuous evolution in the last decades is very significant from a terminological point of view. It is especially from the 1960s, when the use of the computer spread over all human activities thanks to the possibility of using the natural language (Giacomo Ferrari, 2003: 122), that Computer Science becomes a privileged observatory of terminological evolution: in these years, the need of naming rapidly grew within the computer scientists community and the coining of new terms and compound names satisfied this initial need by linking terms to their respective semantic and cognitive value.

Since the advent of the www technologies in the 90's, computing has had a strong impact on modern society offering new opportunities of expansion for future research. In these years the Internet has evolved to such an extent that the important changes in the field of acquisition, storage and transmission of data have provided new resources to the web society, satisfying its needs for constantly updated information. Such information can appear in journals on-line, which were started as such, or of publications originally in paper format and then made available also in electronic form. New forms of publication have emerged which provide the scientific community with free and easy access to research studies, establishing a direct relation between producers and consumers who share knowledge on the web.

The people who, in rapidly increasingly numbers, access the internet to obtain information about Computer Science

and other associated disciplines (i.e. Computational Linguistics) often encounter great problems as much of what they might be looking for can be impossible to find because of inappropriate searches. Different terms used to communicate the same idea can generate linguistic ambiguity, since the same word or phrase can allow for more than one interpretation, thus affecting the information retrieval process. It follows that the queries which are made through these linguistic variations do not always obtain the response looked for, and large amounts of information, although available, do not emerge from the web because the term is not present in the document requested. Access to the semantic contents of a document can become extremely difficult in the case of polysemy (when a word has two or more similar meanings) or of synonymy (when a word means the same as another word).

The word *web* – which abundantly emerges from the corpus - can be taken as an example: Figure 1 shows the use of this term in the scientific documentation belonging to the Information and Communication Technology (ICT) field over the time.

2. PUMA

Data have been extracted from a system called PUMA (Publication Management System), a digital library management system for institutional repositories of technical or scientific documents, either published or self-archived.

PUMA is therefore a software infrastructure, user-focused and service-oriented, developed by the ISTI Institute: the system functionalities manage, for different collections, the whole life cycle of different types of documents, from their submission by authors to their dissemination through web access. The most important PUMA feature is its capability to allow stored content to be reusable for different purposes, so that researchers and librarians can manipulate the stored content to fulfill scientific and administrative issues.

PUMA also constitutes the first step towards creating the Italian network of CNR Institutional repositories, looking at the DRIVER vision, i.e., building an infrastructure that allows European research institutions to share content and functionality. It presently manages 33 CNR institutional repositories that in all contain about 21,000 documents covering different disciplines.

ICT repositories contain about 13,500 documents (the reference corpus of this study already mentioned above) of different types, i.e. published documents – journal papers, books or book chapters, conference papers etc. – and grey literature ones – project deliverables, technical reports, theses and so on. A great part of these documents are Open Access.

2.1 Background Information

ISTI is nowadays entitled to the memory of Alessandro Faedo, who was Rector of the University of Pisa e President of CNR. It is born in 2000 from the fusion of two major Computer Science institutes, IEI (Institute of Information Processing) and CNUCE (National University Centre of Electronic Calculation); IIT was born in 2000 from the fusion between the Institute for Telematics Applications and the Institute of Computational Mathematics and it currently manages the Registry of the Italian Internet domains; ILC was founded by Antonio Zampolli in 1967 as a Department of Computational Linguistics of the CNUCE Institute and it is named Institute of Computational Linguistics in 1978 by initiative of its founder, to whose memory nowadays ILC is entitled.

3. Methodology and data description

Computer Science papers represent the source material chosen for extracting terms. Stand out from titles and abstract of scientific documentation analysed, many subjects of research in the sector of ICT[1].

The archive contains over a million words but only those with a frequency higher than 100 have been selected and analysed: these semantically relevant terms have been channelled in scientific publications over the years and used by the authors in titles, abstracts and titles of their works. Here is a list of both English and Italian terms:

access, accesso, acquisition, addresses, algebra, algorithm, algorithms, algoritmi, allocation, analisi, analysis, annotation, application, applications, architectural, architecture, architectures, artificial, astrophysical, audio, automata, automated, automatic, automatica, automatically, automation, band, bayesian, bit, blind, boundary, cad, calculus, categorization, cep, channel, channels, checking, classical, classification, classifier, cluster, clustering, clusters, codes, coding, collection, collision, communication, communications, complexity, component, components, computation, computational, compute, computer, computers, computing, concept, constraints, context, control, corpora, corpus, database, databases, datasets, dati, debugging, delay, design, designers, development, device, devices, diagnosis, digital, digitale, discovery, distance, domain, domains, earth, elaborazione, electronic, energy, engine, engineering, environment, environments, execution, experiments, extraction, fast, filtering, formalisms, frame, framework, function, functions, fuzzy, gaussian, generation, geo, geographic, gis, graph, graphical, graphics, graphs, grid, grids, hardware, help, heritage, hierarchical, home, ibm, images, imaging, immagini, implementation, informatica, information, informazione, informazioni, infrastructures, input, interaction, interactive, interfaccia, interface, interfaces, internet, interoperability, interpretation, key, knowledge, language, languages, learning, lexical, lexicon, lexicons, libraries, library, linear, linguistic, link, links, location, log, logic, logical, logics, macchina, machine, machines, map, maps, markov, matching, mathematical, matrix, measure, measurements, measures, mechanical, mechanism, mechanisms, memory, metadata, method, methods, microwave, middleware, mining, mobile, mobility, model, modeling, modelling, modello, models, monitoring, multimedia, museum, natural, network, networks, neural, node, nodes, noise, nonlinear, objects, online, ontologies, ontology, operating, operations, orbit, orbital, orbits, output, paradigm, parallel, parameter, path, pattern, patterns, performance, platform, power, practical, privacy, probability, procedures, process, processing, processors, program, programma,

[1] http://www.ict.cnr.it/ :Information Technology (IT), as defined by the Information Technology Association of America (ITAA) is: "the study, design, development, implementation, support or management of computer-based information systems, particularly software applications and computer hardware."In short, IT deals with the use of electronic computers and computer software to convert, store, protect, process, transmit and retrieve information. In this definition, the term "information" can usually be replaced by "data" without loss of meaning. Nowadays it has become popular to broaden the term to explictly include the field of electronic communication so that people tend to use the abbreviation ICT (Information and Communication Technology). Strictly speaking, this name contains some redundancy.

programming, programs, protocol, protocols, prototype, quality, queries, query, radiation, radio, random, range, reasoning, recognition, recovery, reference, remote, rendering, representations, resource, resources, reti, retrieval, risk, run, safety, satellite, satellites, scalable, scanning, schema, scheme, schemes, search, security, semantic, semantics, sequence, sequential, server, sharing, signal, signals, significant, similarity, simulation, sistema, sistemi, sites, software, solar, sound, source, sources, spacecraft, speed, standard, standards, statistical, statistics, step, steps, stochastic, storage, synthesis, system, systems, taken, target, tcp, tdma, technical, technique, techniques, technologies, technology, tecniche, tecnologie, term, terms, tests, text, texts, tool, tools, top, topological, topology, tracking, traffic, training, transmission, tree, understanding, usability, users, utente, validation, vector, video, virtual, vision, web, wide, wireless, xml.

The methodology employed is the following: search and saving of the most common single terms which are the object of this study; extraction of the contexts with the corresponding years; generation of tables according to the chronological use of these terms; creation of charts.

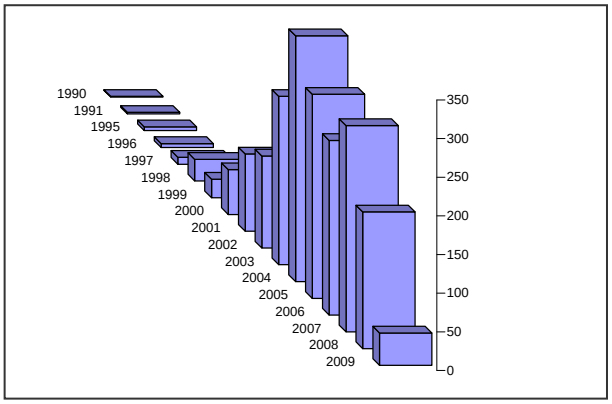

Figure 1: The term *web* over the years

The most frequent cooccurrences in the entire archive have been calculated using the DBT[2] software (see Table 1).

4. A case study

As far as the statistical representation of the research results is concerned, the English term *digital* (*digitale* in Italian) – in singular and plural forms in both languages - appears 1852 times in the archive since 1958. Quoting from Wikipedia: "The word digital comes from the same source as the word digit and digitus (the Latin word for finger), as fingers are used for discrete counting". And also "The word digital is most commonly used in computing and electronics, especially where real-world information is converted to binary numeric form as in digital audio and digital photography".

The processing of concordances on the 13,500 documents of the Archive shows that the term *digital/digitale* is combined with a multitude of words: these co-occurences are showed in Table n. 1 in alphabetical order, with the frequency and the year of the first entry in archive. The table also includes *impronta digitale* (in Italian) with 27 occurrences, comprehensive of those of the English term *fingerprint*[3].

digital 3D	21	2001
digital archive	13	1998
digital certificate	9	2001
digital communication	6	1987
digital computer	12	**1958**
digital content	22	1999
digital conversion	9	1965
digital data	10	1974
digital divide	5	2004
digital document	14	1996
digital filter	10	1974
digital form	17	1996
digital fourier	6	1974
digital image	33	1977
digital information	5	2004
digital library	1158	1995
digital model	39	2001
digital object	22	2006
digital photo	7	2002
digital physics	5	2006
digital preservation	17	2006
digital processing	8	1961
digital repository	15	2006
digital representation	15	1994
digital right	9	2007
digital shape	25	2004
digital signal	26	1975
digital signature	31	2001
digital stream	6	2006
digital system	34	1966
digital techniques	7	1995
digital terrain	13	1981
digital TV	21	2003
digital video	12	2005
impronta digitale	27	1968

Table 1: List of co-occurrences

The dating of the first documentary evidence of the entry *digitale* in Italy goes back to 1961[4] while in the United Kingdom the entry dates back to 1938[5]. Since then the use of this adjective is continously evolving and year after year the need of coining new terms – mainly related to technological developments – brings about the creation of new couples of terms. The search engine Yahoo, for example, provides 213.000.000 results for the word *digitale* and 3.640.000.000 results for *digital* [6].

We have split the above data into two graphs (see Figures 2 and 3) because of the great difference between the frequency of *Digital Library* and the frequency of the other co-occurrences.

The concordances extracted from the data archive witness the combination from time to time and from year to year of the adjective *digital* with other nouns such as *archive, image, photo, signal, signature, TV, rights, repository,*

[2] DBT (Data Base Testuale) CNR patent by Eugenio Picchi.

[3] In our archive this term includes the concept of digital image.

[4] Word *digitale*, il Sabatini Coletti, Dizionario della lingua italiana. Milano, Rizzoli Larousse, 2006, p. 740.

[5] Word *digital*, Historical Thesaurus of English.

[6] 2010-02-25, h 12:15.

terrain, etc. These combinations stress the spreading of Computer Science to important topics of research.

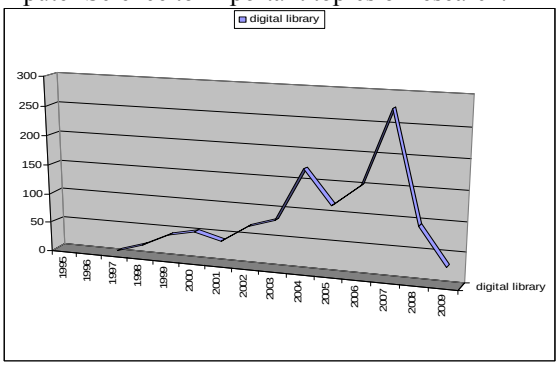

Figure 2: *Digital Library* over the time

The thematic content of the adjective is given by the noun combined with it: this is the case, for example, of the combination *digital signal* which testifies to a technological progress. In many cases, terms develop over the time, establish themselves and then transform themselves into something new but still prove the scientific activity reflected in the documentation.

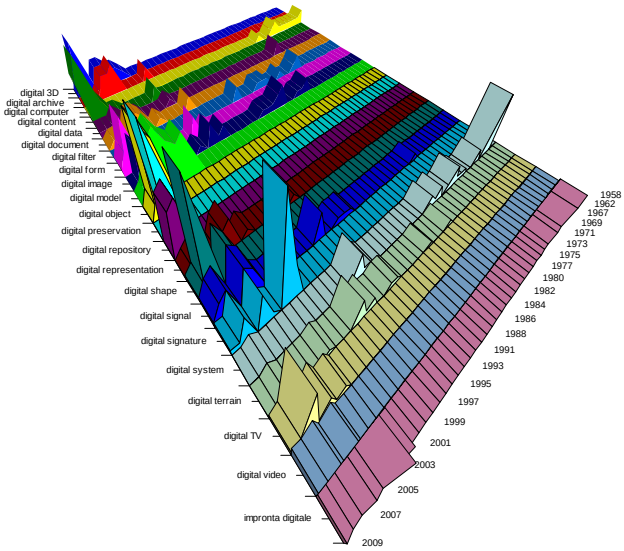

Figure 3: *Digital* co-occurrences

5. Conclusion

Through space and time, terminology is a way for acquainting with the culture of a discipline.

As a conclusion, it is very significant what Teresa Cabré says about the relationship between documentation and terminology: "While terminology is at the service for documentation, documentation is also at the service of terminology. Technical documents are the basis for terminological work. Terminologists do not invent designations for concepts in a previously established system, but rather identify and collect terms that specialists use in documents. Terminologists use documents in order to acquire knowledge of the subject and its conceptual structure , to search for the term used in a subject field, to confirm the quality of the primary data they have collected, to study the data from the various viewpoints represented by different documents" (Cabré, 1999: 51).

6. References

Biagioni S., Carlesi C., Romano G. A., Giannini S., Maggi R. (2007). PUMA & MetaPub: open access to Italian CNR repositories in the perspective of the European Digital Repository Infrastructure. In Dominic Farace (ed.), *9th International Conference on Grey Literature.* Antwerp. Proceedings. Amsterdam, TextRelease.

Bird, S., Dale, R., Dorr, B.J., Gibson, B., Joseph, M.T., Kan, M.-Y., Lee, D., Powley, B., Radev, D.R., Tan, Y.F. (2008). A Reference Dataset for Bibliographic Research in Computational Linguistics. In Proceedings of the *Sixth International Language Resources and Evaluation*. European Language Resources Association, Paris.

Cabré M.T. (1999) *Terminology: theory, methods, and applications*, Edited by Juan C. Sager, Amsterdam /Philadelphie, John Benjamins.

Cabré M.T. (2000). Terminologie et linguistique: la théorie des portes. *Terminologies nouvelles. Terminologie et diversité culturelle,* 2, pp.10--15.

Cabré M.T. (2008). Realidad, cognición y lenguaje : la poliedricidad como principio. Atti del Convegno Nazionale Ass.I.Term, Università della Calabria. *AIDA Informazion*i, 26(1/2), pp. 11--24.

Marzi C., Pardelli G., Sassi M. (2009). Grey Literature and Computational Limguistics: From Paper to Net. In D. Farace and J. Frantzen (eds.), *11th International Conference on Grey Literature*, Washington. Proceedings. Amsterdam, TextRelease, pp. 81--84.

Pardelli G., Sassi M., Goggi S., Orsolini P. (2009). Computational Linguistics Terminology. In *Actas XI Simposio Internacional de Comunicación Social. Centro de Lingüística Aplicada*, Ministero de Ciencia, Tecnología y Medio Ambiente, Santiago de Cuba, pp. 303--307.

Pritchard, D. (2008). Working Papers, Open Access, and Cyber-infrastructure in Classical Studies. *Literary and Linguistic Computing*, 23 (2), pp. 149--162.

Sassi, M., Pardelli, G., Goggi, S. (2009). Terminology Extraction from the Web. In Z. Vetulani (ed.), *Proceedings of 4rd Language & Technology Conference*. Fundacja Uniwersytetu im A. Mickiewicza, Poznań, pp. 417--421.

Temmerman R., (2000). *Towards New Ways of Terminology Description. The sociocognitive approach.* Amsterdam /Philadelphie, John Benjamins.

Ruimy N., Bozzi A., Pardelli G. (2009). Modèle lexical pour un thésaurus-lexique électronique de la terminologie saussurienne. *Séminaire international Publier les manuscrits de Ferdinand de Saussure*, Arcavacata, Università della Calabria, 1-3 ottobre 2009.

Zampolli, A. (1973). Humanities Computing in Italy. *Computers and the Humanities*, VII (6), pp. 343--360.