International Conference on International Knowledge Circulation,  
Mexico City, October 9-11, 2012

Computer supporting research on Diaspora Knowledge Networks

William Turner¹, Mathilde de Saint Lege², Jorge Garcia Flores¹, Zhao Yue², Jean-Baptiste Meyer², Pierre Zweigenbaum¹

1. LIMSI-CNRS, Building 508, Université de Paris Sud, 91400 Orsay, France.
2. LPED-IRD, Université de Provence, 3, place Victor Hugo 13331 Marseille, cedex 03 France
3. MoDyCo, Université Paris Ouest, 200 Avenue de la République, 92001 Nanterre la Défense.

Summary

National governments are reaching out to talented members of their diasporas in attempts to engage them in projects for home country development. These “reach out programs” are organized around three activities. First, statistical sources of information are generally used to determine the institutional and geographical location of talented people who have gone overseas. Population censuses, population registers, labour force surveys, administrative data, specific surveys and case studies are used to measure migration flows, however, these are not entirely satisfactory when it comes to monitoring a constantly changing situation. In this paper we will report on our work aimed at datamining the Web in order to update on a steady, on going basis the understanding of mobility trends affecting national science systems. Describing these trends is the second problem. We’ve dealt with it at two levels. The first is to carry out what we call a “Mobility Trace Classification” task in order to help sociologists answer the following types of questions: Among the Latin-American authors who published scientific articles about biotechnology during 2011, how many of them are living abroad? And how many of them have studied in foreign universities before coming back to work in their home countries? The second level of analysis derives from the assumption that different forms of mobility – be it linear, when a person leaves his or her home country and establishes residence permanently abroad, or circular, when he or she comes back home to live after periods spent outside the country – will have an impact upon the ways in which intellectual, human and material resources for scientific work are accessed, captured and mobilized. People on the move build networks of cognitive, institutional and spatial relationships which we are attempting to map out in order to help organize national reach out programs. The third question is to know how to make the best use of these talent mobility maps for mobilizing the support of Diaspora Knowledge Networks for home country development. In order to answer it, measures for comparing what we know of local research systems and what we show on the talent mobility maps are being produced.

This paper is organized in four sections. After introducing our approach to computer supporting Diaspora Knowledge Networks using techniques for datamining the Web, in the second section we will present the UNOPORUNO system we’re developing and first results on its performance in carrying out the Mobility Trace Classification task. The third section of the paper will examine technical issues faced in using the material extracted from the Web for building talent mobility maps. Finally, we will discuss examples showing how these maps can be used to reach beyond the borders of national science systems in order to access the cognitive and human resources of research institutions in foreign countries.
1.) Introduction: policy tools for monitoring international talent mobility

Diaspora Knowledge Networks (DKNs) are self-organizing social structures that serve to identify, capture and mobilise skills and knowledge abroad for home country development (Turner, 2006). Empirical evidence suggests that in some cases self-organization dynamics can be sub-optimal in terms of organizing this mobilization. One example concerns the fact that junior scientists generally exploit their supervisor’s network of personal and professional relationships when starting out on their own career (Melin, 2004). As a result, mobility channels are installed and reinforced over time locking research institutions from developing countries into fairly stable patterns of communication and exchange with a limited number of Western science centres (Mahroum, 2000). These mobility channels show little plasticity: they tend to reproduce and consolidate existing socio-cognitive relationships rather than call into question their relevancy for home country development (Woolley et al, 2008; Kreimer et Zabala, (2008); Melin 2004).

Our efforts to computer support DKNs are anchored in this observation: what type of policy tools are needed in order to monitor patterns of international talent mobility and critically analyze the impact of these patterns on the scientific output of national research systems? A policy tool is an information processing device which helps actors (decision-makers, stakeholders, the general public) evaluate the degree to which they are locked into a specific way of doing things (Turner et al, 2009). Maps, for example, are a policy device. They serve to visualise the social dynamics of a collective practice by identifying a population and using different measures to illustrate the ties which exist between members of that population. Among the measures the most frequently used are the density of ties and the centrality of an element’s position in a network – the element being a person when the study is designed to explore social relationships, an institution when the goal is to study institutional ties or a concept if the goal is to map out cognitive relationships. Our interest for a given person, institution or concept is mise en scène by the fact of seeing that person, institution or concept as central (or peripheral) in the overall structure of the map; or as being densely or weakly connected with the other elements of the general picture. And this, in theory, is what encourages discussion of alternative ways of acting.

Lists are another policy tool for making explicit the categories which often implicitly underlie social practices. As Goody has shown, building an inventory and constantly updating it implies an on-going discussion of what is similar and what isn’t when putting elements on that list (Goody, 1979). For example, the subjects examined in connection with clinical medicine in Latin America are not necessarily the same as those treated in a European medical context. Different forms of knowledge and skills come into play which are both a justification for organizing talent mobility and a factor complicating its usefulness and success for home country development.

Finally, computers are increasingly able to draw their own conclusions from the data processing they are programmed to accomplish, and make recommendations on the basis of those conclusions. These recommendations are designed to alert people to the need for acting in new ways; to show them that there is a need for breaking out of the “lock-in dynamics” structuring their collective practices.

This paper aims at showing how computer generated maps, lists and recommendations can be used as a means of collectively building the story which will help governments engage the talented members of their diasporas in home country development projects.
2.) Building lists for telling the story of people on the move

From a computer science point of view, success in carrying out the Mobility Trace Classification Task depends upon developing algorithms which optimize list management techniques. The starting point is a list of bibliographical records downloaded from the ISI Web of Knowledge\(^1\) concerning publications produced by Uruguayan researchers over a 10 year period from 2001-2010. For this paper, 2269 records were selected, providing us with an initial list of 8474 co-authors (roughly 4 per paper) who were considered as potential talent mobility candidates. Of these, 5913 names were concretely processed after elimination of spelling errors and name repetitions. UNOPORUNO sent 74137 queries to Google who returned to the system 1545458 snippets in response to its questions. The following table shows a list of three snippets that UNOPORUNO selected as being particularly useful for deciding upon the mobility status of Alberto E. Pereda, one of the authors in the downloaded ISI file.

![Snippet](source)

### Table 1: The UNOPORUNO user interface

The usefulness of the snippets selected by UNOPORUNO is validated by a sociologist. The system’s interface is shown in Table 1. The upper part of the screen contains the name of the person, together with information extracted from the ISI database on his geographical location, institutional affiliation and area of subject interest. It also contains the system’s conclusion as to the person’s mobility status (“status probable”) and the decision taken by the sociologist with respect to the real status of the person (“status real”).

The lower part of the screen contains three Google generated snippets. As can be seen from Table 1, a snippet provides a succinct summary of the information contained on the page referenced by the snippet. The important point here, and the one we will come back to in a moment, is the assumption that the value of a snippet lies in the way specific features are combined in the body of the snippet. We’ve been working at identifying those features; they are listed on the left-hand side of the screen under the heading “caracteristicas” and correspond to the categories listed in table 2 below. Each category is composed of a list of linguistic markers. For example, we use:

\[^1\]http://www.isiwebofknowledge.com/
- A multilingual database containing a list of 4376 entries to identify world cities having populations over 100000 residents;
- A multilingual list of countries containing 516 records, etc.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD thesis</td>
<td>regex</td>
<td>The snippet links to a PhD thesis</td>
<td>bool</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>gazet</td>
<td>The snippet links to a LinkedIn Web page</td>
<td>bool</td>
</tr>
<tr>
<td>Publication</td>
<td>gazet</td>
<td>The snippet links to a scientific publication</td>
<td>bool</td>
</tr>
<tr>
<td>e-mail</td>
<td>regex</td>
<td>The snippet contains an email</td>
<td>bool</td>
</tr>
<tr>
<td>Non Latin-American nationality</td>
<td>gazet</td>
<td>The snippet contains a nationality from a non Latin-American country</td>
<td>bool</td>
</tr>
<tr>
<td>Latin-American nationality</td>
<td>gazet</td>
<td>The snippet contains a nationality from a Latin-American country</td>
<td>bool</td>
</tr>
<tr>
<td>Person name found in URL</td>
<td>regex</td>
<td>Personal first or last name in the http address</td>
<td>bool</td>
</tr>
<tr>
<td>CV</td>
<td>regex</td>
<td>The snippets links to a CV</td>
<td>bool</td>
</tr>
<tr>
<td>Profession</td>
<td>regex</td>
<td>The snippet contains a profession name</td>
<td>bool</td>
</tr>
<tr>
<td>Degree</td>
<td>regex</td>
<td>The snippet contains academic information</td>
<td>bool</td>
</tr>
<tr>
<td>Biographical sentence acronym</td>
<td>gazet</td>
<td>The snippet contains a biographical sentence acronym</td>
<td>bool</td>
</tr>
<tr>
<td>Organization</td>
<td>gazet</td>
<td>The snippet contains an organization acronym</td>
<td>bool</td>
</tr>
<tr>
<td>City &amp; region</td>
<td>gazet</td>
<td>The snippet contains a city or region name</td>
<td>bool</td>
</tr>
<tr>
<td>Country</td>
<td>gazet</td>
<td>The snippet contains a country name</td>
<td>bool</td>
</tr>
<tr>
<td>Organization</td>
<td>regex</td>
<td>The snippet contains an organization name</td>
<td>bool</td>
</tr>
<tr>
<td>Feature count</td>
<td>-</td>
<td>Number of features found in the snippet</td>
<td>int</td>
</tr>
</tbody>
</table>

Table 2: Semantic features for snippet analysis

Sociologists use the right part of the screen shown in Table 1 to evaluate the usefulness of a snippet. They are asked if the value of the snippet is weak (débil) because it doesn’t allow direct access to relevant information or strong (fuente) because, to the contrary, a simple click on the link in the snippet provides immediate access to appropriate information (a CV, a faculty Web page, a social network profile, etc.). Or they can choose not to annotate the snippet leaving the box blank and indicating that the snippet isn’t of use for deciding upon the mobility status of the person. By observing the decisions taken by sociologist, the machine learns to recognize the combination of features which are likely to be the most relevant for solving what has been called the “needle in the haystack problem”. This problem derives from the fact that sociologists don’t want to spend an enormous amount of time wading through a huge amount of data in order to find the appropriate information for carrying out the mobility classification task.

The figures above provide us with a starting point for discussion the needle in the haystack question. As we saw, when different types of information on geographical location, institutional affiliation and subject area interests are combined with a person’s name to produce Google queries, UNOPORUNO submits on an average of about 12 to 15 multilingual queries per person and, for each query, Google returns approximately 21 snippets. In theory, then, for each mobility candidate, a sociologist would have to view about 275 Web pages in order to process all the information potentially useful for carrying out the mobility trace.
classification task. This is clearly impossible; the task would be much too time consuming to think that sociologists would be willing to engage in it.

UNOPORUNO offers a semi-automated solution to this problem by using semantic filters and statistical classifiers in order to rank snippets according to the relevancy of their feature combination configurations for carrying out the mobility classification task. As shown in Table 1, the user interface opens onto a screen where only the Top 5 snippets in this ranking are presented for evaluation by the sociologist. In other words, with UNOPORUNO, sociologists are expected to find relevant information with only 5 page views which is only 2% of the data she would have had to wade through without UNOPORUNO (275 pages). Finally, if a sociologist does not gain access to relevant data using the Top 5 snippets, she can look at all the data (the “todos” button shown in Table 1) collected by Google on the cited person.

The question, then, is to know if the two steps of the UNOPORUNO pipeline are working correctly. The first step concerns the application of our natural language processing tools for detecting in snippets the presence or absence of the different semantic features shown in Table 2. The second is the statistical classifier used to produce the snippet relevancy ranking justifying our decision to present only 2% of the available data to sociologists in order to avoid the needle in the haystack problem. We’ve answered these questions technically in a recent paper (Garcia Flores, 2012) and will not describe here the work we carried out to build scientifically relevant training and testing corpora; to test different statistical classifiers; to improve the quality of our feature analysis; and to apply appropriate accuracy, recall and precision measures.

Here, however, are the results: in roughly 80% of the cases, viewing the Web pages referenced by the Top 5 ranked snippets enabled a sociologist to successfully carry out the mobility classification task. 7 sociologists participated in the experience, so we were able to compare the snippets that they individually annotated as being useful and strong (“fuente”); useful and weak (“débil”); or didn’t annotate because they were irrelevant. We found a moderately high level of inter-evaluator agreement ($\kappa$) of around 60%. Finally, we found that performance levels are modified when the following semantic features in Table 2 are not included in the test runs: PhD thesis, Publication, Organization Acronym, City, Feature Count, Profession and City. However, the absence of other categories such as Organisation, Academic Degree and Country, doesn’t affect test run performances. This finding is currently being explored because it implies that our representation of useful categories for opening a window onto the Web might have to be changed.

What, then, is the general conclusion of the discussion in this section of the paper? We’ve shown that with UNOPORUNO we can build a potentially relevant corpus of snippets in response to queries submitted to a search engine; that these snippets can be automatically classed and ranked in a way which drastically limits to only 2% the number of page viewings needed to take a decision on the mobility status of a researcher, and that the decision itself is based on a semantic understanding derived from the snippets of what is contained on the Web pages they reference. Our claim is consequently that snippets can be used as a surrogate for these documents. In order to data-mine the Web we don’t have to access Web pages; we can use snippets. This point is far from evident when we look at the snippets in Table 1. We see that they contain a very small number of words which legitimately raises the question of their information value for studying talent mobility. But what we’ve shown in this section, is that their information value is real: snippets can be used directly for producing maps to show the
impact of mobility on local research systems; the results are meaningful. This is what we will show now.

3.) Using snippets to build maps for critically analysing the impact of talent mobility on national science systems

The strategy of using snippets as a surrogate for a Web page analysis can be evaluated in terms of the solutions which it offers for describing the spatial, institutional and cognitive dynamics of talent mobility. More concretely, three questions have to be addressed when mapping out these dynamics. The first is to know how to frame the picture: what should go onto the map and what should be left out? We saw in the preceding section that only two percent of the snippets returned in response to a question are theoretically sufficient for building a relevant map. But what would happen if, instead of using the Top 5 snippet selection, we used the Top 20, or the Top 50 or the Top100? When snippets with low relevancy ratings are used to build maps, two opposite results can be envisaged: either they serve to enhance our understanding of talent mobility by showing what is increasingly peripheral, but nevertheless relevant, to framing the phenomenon we want to study; or we clutter the picture up with a wealth of insignificant details making it very complicated to apprehend. In order to study the impact of relevancy thresholds on framing a picture of geographical mobility, we worked with two files: one constructed using the Top 5 classification of snippets for 25 members of Uruguay’s talented diaspora; the other constructed for the same researchers, but this time using a Top 20 threshold. The maps show relationships between the countries which are identified in the snippet files corresponding to the two thresholds. They will be compared in the next section of this paper.

The second question concerns the meaning of links which are represented on a map. These links have to be meaningful for those people that will use the map. As we said at the outset of our paper, the users of our maps are managers of “reach out programs”. We are trying to help them by working on the assumption that researchers working in the national science system will have a different set of institutional, cognitive and spatial relationships to those that have left the system and are working abroad, and that it is precisely in attempts to play upon those differences and complementarities that the construction and consolidation of Diaspora Knowledge Networks makes sense. But how do you show on a map, the opportunities for knitting together local and diaspora knowledge networks? The minimum requirement is that the representation of these networks is comparable. And this is guaranteed by extracting the nodes and the edges of the graph from a snippet dataset. On the one hand, the semantic feature analysis described in the preceding section ensures that the nodes designating institutions, concepts and countries are common to the two corpora and, on the other hand, because a snippet is a surrogate for a Web page, we know that the co-presence of words in the snippet is the expression of a semantic relationship established intentionally in the text of the Web page. And its this assumption of intentionality which enables us to consider the relationship of two words in a snippet as significant. In the following section, we will illustrate this point by comparing the network of country associations in which the local Uruguayan science system operates to a representation of the country associations in which Diaspora science is produced.

The third point concerns the conditions for making these comparisons. Because snippets are surrogates for documents, we consider that it is legitimate to use the Calliope suite of tools, which was designed for use in a text mining context, to assist us in establishing the conditions of this comparison. What is involved? Calliope was designed to study language as an
expression of social behaviour (de Saint Leger, M. 1997). It, too, focuses on the co-presence of words in a text and interprets that relationship as being an invitation to collectively consider the opportunity of connecting two countries, two institutions, two concepts, a country and an institution, a concept and a place, etc. The problem is the huge variety of words encountered when analyzing a text and the fact that not all words serve as markers of an intention to link specific resources together. Calliope provides us with a platform for normalizing vocabulary in order to deal with this question, as can be seen in Table 3.

Table 3: The Calliope platform for vocabulary standardization

All the words that Calliope automatically extracted from the snippet file appear in the left hand column of Table 3. Various options can be used for visualizing this list: frequency, alphabetical order, validated terms, etc. Above, on the upper line of the screen shot, different options are shown for working with lexicons: they can be imported as a list of authority in order to validate the relevancy of a vocabulary extraction; exported for integration into a constantly, updated list of authority; or one can create his or her own terminology depending upon the his lexicon management needs. On the right hand side of the screen shot, under the heading “lexique de report” the list of elements shown with “Institut Pasteur” illustrates the decisions which are being taken to normalize how this institutional resource is being represented for the machine. Finally, in the upper right hand corner is a list of all the terms eliminated from the vocabulary.

A lot of fine tuning has to be done before the spatial, institutional and cognitive dynamics of talent mobility can be suitably visualized. We’ve shown in this section that the fact of considering snippets as surrogates for the documents to which they refer is useful for this purpose. It has allowed us to develop tools for empirically dealing with the problem of framing the picture of what is going on in the talent mobility sector; for identifying different configurations of institutional, conceptual and geographical resources; and for doing the technical maintenance of these network management tools.
We will now turn to a presentation of the results obtained when applying our tools to different snippet datasets. These results have been produced for demonstrative purposes. They concern very small data sets for only 25 researchers manually classed as local scientists, and for 25 classed as diaspora scientists. We produced data on circular scientists, but have not had the time to fully process them yet.

4.) Questions for decision-makers

The Calliope screen print in table 4 shows the cities and countries appearing in the Top 20 snippet set for researchers manually classed as “locals”, that is, for those people who have only passed short periods of time abroad (less than one year). The screen is divided into two parts: an upper light green part which shows on the right vertical axis the countries and cities cited in the snippets and a lower darker green area which falls under a black line. The goal of this diagram is to show the words structuring a collective activity, the black line being a threshold: above the threshold the cities and countries are considered, according to the algorithm, as a backdrop to the science which is produced locally in Uruguay; below the threshold their influence on the science produced in Uruguay is not significant. Australia was intentionally left on the screen shot in order to make the point that its position on the threshold can be interpreted as confirmation of the fact that it is not a country of reference for the local Uruguayan science network. What is striking on the map is the fact that none of the science centres in the United States and Western Europe (with the exception of Granada) appear on the map.

Table 4 : A representation of the country associations in which Uruguayan local science is being produced (Top 20 Threshold)

Table 5 contrasts with table 4 by the fact that Uruguayan Diaspora science is being produced in a context which is clearly oriented towards Western Europe and the United States. The other notable feature of the screen shot is that the focus has changed. The high points are now in the middle of Table 5 instead of being on the right hand side. The explanation is simple: for this screen shot, three snippet
files were analysed corresponding to 25 circular, 25 diaspora and 25 local scientists. The results are compared using the threshold: the cities and countries structuring diaspora science are shown in the middle of the screen (above the threshold); these countries don’t, however, contribute to the structure of local or circular science networks (below the threshold).

Table 5: A representation of the country associations in which Uruguayan Diaspora science is being produced (Top 20 Threshold)

Table 6: A representation of regional patterns of talent mobility (Top 20 Threshold)
Table 6 uses an open source graph visualisation and manipulation software called GEPHI to map out the same Calliope data as in table 5. The goal in producing Table 6 was to show the countries cited in the Diaspora snippet set for each of the three regional mobility destinations; Europe, the United States and Latin America. The vocabulary serving to construct table 6 was also normalized and the map was build around country rather than city names. Table 7 is the same as table 6 but was produced using a Top 5 threshold instead of a TOP 20 threshold. The change in representations is spectacular. There are 3 times fewer nodes obtained in exploiting the Top 5 dataset, and the United States disappears altogether as a place structuring Uruguayan diaspora research.

Table 7 : A representation of regional patterns of talent mobility (Top 5 Threshold)

Once again the results presented in this section serve to make methodological points and should not be considered as a window upon the reality of Uruguayan research. Our sample is too small to draw any substantive conclusions. Nevertheless, we wanted to show that we are starting to produce policy tools that can be of use in reaching out to members of the diaspora and in positioning them with respect to the needs of a local research system. These tools still need a lot of fine-tuning, but a case for computer supporting the formation and consolidation of Diaspora Knowledge Networks can now clearly be made.

Bibliography

1. de Saint Leger, M. 1997 ; Modélisation des flux d’information scientifique et technique pour un suivi de l’évolution des domaines de connaissances ; Thèse de Doctorat d'Université, CERESI/CNRS-CNAM ; 1997.

Biographies:

William Turner

William Turner is a Senior Research Engineer and member of the French National Research Centre (CNRS). He currently heads the social informatics research team at the Computer Science Lab for Engineering and the Mechanical Sciences (LIMSI). He started his career at the Paris School of Mines working in the Centre for the Sociological Study of Innovation, before moving into information science and taking the head of programs aimed at computer supporting distributed collective practices in both science and industry. He has served as an expert to the National Science Foundation in the United States, the European Commission, the French National Agency for Innovation and the French Institute for Research on Development. He has published over one hundred articles in scientific journals ranging from computer science to sociology and focusing mainly on the cognitive management of social networks, and the impact of this management on policy making.

Mathilde de Saint Leger

Mathilde de Saint Leger has recently joined MoDyCo (the University of Paris-West, Nanterre, Dynamic Modeling and Corpus Laboratory) where she is a research engineer responsible for the statistical applications of this Computer Science Lab. She has a PhD in Information Science and a long experience in bibliometrics, econometric intelligence and the mathematical modeling of information flows. She created and developed the textmining software, Calliope, and has served as project manager for both the Scientific and Technical Information Directorate of the National Scientific Research Centre (CNRS) and the National Agricultural Research Institute (INRA) in France.

Jorge J. Garcia Flores

Jorge J. Garcia Flores (Mexicali, 1972) is a Computational Linguistics researcher. He studied Computer Science Engineering at Monterrey Tech (Mexico). He has a PhD in Computational Linguistics from Paris-Sorbonne University. His research subjects are natural language processing, semantic filtering, automatic summarization, web people search and NLP for
requirements engineering. He has been a post-doctoral fellow at the French Atomic Energy Comission and at LIMSI-CNRS labs in France.

Zhao Yue

ZHAO Yue is an intern at the Institute of Research for Development, currently working with Mr. Jean-Baptiste Meyer on the CIDESAL projet. He got his Bachelor of Arts degree at the Paul Valéry University in Montpellier, France, and is now in the second year of his Master’s Degree in the Social Science Department of the Paul Valéry University in Montpellier.

Jean-Baptiste Meyer

Jean-Baptiste Meyer is senior researcher at the Institute of Research for Development. He has run research and higher education programmes with the National University of Colombia, the University of Cape Town and the Latin American Faculty of Social Sciences in Buenos Aires. His works include: *El nuevo nomadismo científico: la perspectiva latinoamericana* (ESAP 1998), *Scientific Diasporas* (IRD editions, 2003), *La société des savoirs: trompe l’oeil ou perspectives* (Harmattan 2006), *A sociology of diaspora knowledge networks* (forthcoming 2012). He is currently coordinating the CIDESAL european research and development project, on diaspora incubators, developing new methods and instruments for global mobility understanding and management. [http://www.observatoriodiasporas.com](http://www.observatoriodiasporas.com)

Pierre Zweigenbaum

Pierre Zweigenbaum is Senior Researcher and head of the Natural Language Processing group at LIMSI, a unit of CNRS, the French National Council for Scientific Research, and is Invited Professor at the National Institute of Oriental Languages and Civilizations, where he teaches Natural Language Engineering. He received a PhD in Computer science from Telecom ParisTech in 1985, and the “Habilitation à diriger des Recherches” in 1998 from Paris-North University. Before CNRS, he worked from 1984 to 2006 in the Department of Medical Informatics of Assistance Publique - Hôpitaux de Paris, then Research Mission for Medical Information Sciences and Technology, and has been affiliated with INSERM (research units U194 and U729, Knowledge Engineering in Health).

His work has focused on Natural Language Processing (NLP) in Medicine, from morphology to syntax to knowledge representation. He was coordinator of the European project MENELAS, on the analysis of natural language patient discharge summaries (1992-1995), and team leader in two other European projects (NLPAD and DOME) on medical language processing. He coordinated the French project UMLF on the development of a unified medical lexicon for French. More recently he was team leader for LIMSI in ANR-funded projects Akenaton, InterSTIS, C-Mantic and currently Accordys. He is co-coordinator of the Information Extraction task in the French-German Quaero programme.

Pierre Zweigenbaum has taught NLP in Language Engineering, Medical Informatics and Artificial Intelligence curricula. He has been chief editor of the French journal Traitement Automatique des Langues, is a member of the editorial board of the Terminology journal and of the Revue d'Intelligence Artificielle, and reviewed papers for international journals and conferences in NLP and in Medical Informatics. He chaired or organized national and international conferences and workshops on Natural Language Processing, Computational Terminology, Text Mining, and Comparable Corpora.
He was chairman of the Natural Language Processing working group of the American Medical Informatics Association, Vice-President of the French NLP Society (ATALA), Secretary General of the French Association for Medical Informatics (AIM), board member for the French Association for Artificial Intelligence, coordinator of the “language” section of CNRS-sponsored “GDR Information - Interaction - Intelligence” research federation, and moderator of the international LN electronic bulletin on NLP.

He is the author of over 200 refereed publications (http://www.limsi.fr/~pz/biblio-pierre-pardate/).