Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 69, No. 2 (2006) 323–345

The unification of institutional addresses applying parametrized finite-state graphs (P-FSG)

CARMEN GALVEZ, FÉLIX MOYA-ANEGÓN

Scimago Research Group, Department of Information Science,University of Granada, Granada (Spain)

We propose a semi-automatic method based on finite-state techniques for the unification of corporate source data, with potential applications for bibliometric purposes. Bibliographic and citation databases have a well-known problem of inconsistency in the data at micro-level and meso-level, affecting the quality of bibliometric searches and the evaluation of research performance. The unification method applies parametrized finite-state graphs (P-FSG) and involves three stages: (1) breaking of corporate source data in independent units of analysis; (2) creation of binary matrices; and (3) drawing finite-state graphs. This procedure was tested on university departmental addresses, downloaded from the ISI Web of Science. Evaluation was in terms of an adaptation of the measures of precision and recall. The results demonstrate the usefulness of this approach, though it requires some human processing.

Introduction

Bibliographic databases constitute a source of problems when used for informetric and bibliometric purposes. One major difficulty lies in the errors and the lack of consistency in data. Moreover, when these databases are used as potential resources for building scientometric indicators, another series of technical obstacles arises in the combination, or "hyphenation", of databases owing to different standards in abbreviation, spelling and transliteration (BRAUN et al., 1995). The quality control of

Received January 17, 2006

Address for correspondence:
CARMEN GALVEZ
Scimago Research Group, Department of Information Science, University of Granada
18071, Granada, Spain
E-mail: cgalvez@ugr.es

0138–9130/US $ 20.00

data, both within and across databases, is a necessary issue, often addressed yet to date unsolved (SHER et al., 1966; GARFIELD, 1979; 1983a, b; WILLIAMS & LANNOM, 1981; PITERNICK, 1982; STEFANIAK, 1987; MOED & VRIENS, 1989; PAO, 1989; RICE et al., 1989; CRONIN & SNYDER, 1997; INGWERSEN & CHRISTENSEN, 1997; GILES et al., 1998; HOOD & WILSON, 2003). In a secondary use of these databases, in bibliometric analysis, again databases of bibliometric information are created, their construction involving some or all of the following steps: information gathering, information processing, data standardization, and codification. Both for the database producers and for those downloading the data in quantitative studies, the lack of standardization and related errors can lead to the loss of information. As a result, the non-unification of data makes it necessary to perform offline correction to ensure the rigor of research evaluation, largely dependent on the quality of bibliometric data.

In general, the importance of standardizing corporate source data resides in the increasing weight of studies about research centred on institutional domains. More specifically, the inaccuracies in the names of organizations in scientific publications (such as spelling variants, typographic errors, the incorrect use of capital letters, use of initials, abbreviations or transliterations) may distort the results of bibliometric analyses, and the unification of data calls for careful and costly manual cleaning-up processes. While this is by no means a new problem, its incidence shows a spiralling trend, as the citations to scientific articles, stored in databases, can now be used by science policy-makers, and the consequences of citation analysis errors can be great. The present paper aims to contribute to the standardization of corporate source data, in view of the affluence of repercussions at the micro-level of institutional affiliation, such as collaboration indicators, delimitation of scientific fields and dynamic aspects of scientific research.

Collaboration indicators. Affiliation data are commonly used as indicators for scientific collaboration, based on the analysis of all addresses in papers published by a research unit. The type of collaboration may be inter-departmental (between two departments), inter-institutional (collaboration involving institutions of a single country) or international (with at least one foreign address). In the case of scientific collaboration, the structure of the addresses makes it possible to study co-authorships – a scientific document is institutionally co-authored if it has more than one author address, suggesting that the authors come from various institutions, department or other kinds of units – using main organizations, cities and countries as the unit of investigation (MELIN & PERSSON, 1996). Bibliometric studies of scientific collaboration, either within or among research groups or countries, are increasingly frequent (RINIA et al., 1993; HERBERTZ & MÜLLER-HILL, 1995; VAN DEN BERGHE et al., 1998; MOED, 2000). Thus, "for assessing international cooperation connections, unified addresses relating to institutes, cities, and countries are extremely important" (DE BRUIN & MOED, 1990, p. 76).

Delimitation of scientific fields. The institutional affiliation data can be used for the delimitation of scientific fields and subfields, building maps through the co-occurrences of cognitive words in the addresses (DE BRUIN & MOED, 1993). Problems stem from the incongruencies between the department name and the designation of fields and subfields, or the lack of correspondence to the actual sites of research of authors, and may require complementary data, for instance classifying scientific publications by field of research (BOURKE & BUTLER, 1998).

Dynamic aspects of scientific research. Evaluative scientometrics, a subarea of bibliometrics, has developed arrays of indicators that can be used to describe the research performance of the organizations at different levels of aggregation (CARPENTER et al., 1988; MOED & VAN RAAN, 1988). The assessment of organizations provides new possibilities for the utilization of the address field in a picture of the dynamic aspects of scientific research, such as the mapping of science and network analysis (SHRUM & MULLINS, 1988; NOYONS et al., 1999; MÄHLCK & PERSSON, 2000). At the same time, the development of new analytical tools based on the combined use of methodologies – multivariate analysis, artificial neuronal networks, and techniques based on network analysis – makes it possible to represent research in a given scientific domain in the form of dynamic maps and science atlases. Recent work on the application of such methods to the representation of research in institutional domains (MOYA-ANEGÓN et al., 2003; 2004) underlines the need to standardize corporate source data.

The errors and inconsistencies in addresses can affect indeed not only the study of co-authorships, the evaluation of international cooperation connections, and the delimitation of scientific fields, but also the visibility of institutions and the ranking of research organizations. Discussion of the non-standardization of institutional affiliations can be found in ANDERSON et al. (1988), LEYDESDORFF (1988), DE BRUIN & MOED (1990), BOURKE & BUTLER (1996), and VAN RAAN (2005). Bibliometric institutes such as the Leiden Center for Science and Technology Studies (CWTS) have expressed interest in the standardization of this type of data. We therefore propose a formalism that will make possible the unification of address data by means of a partially automated procedure.

Lack of consistency of corporate source data

Scientific publications habitually contain interrelated data as to the institutional affiliation of authors, in general comprising four parts: the main organization, the department of that organization, the city and the country (MELIN & PERSSON, 1996). The critical problem of corporate source data derives from variants of the name of a

given university, department or research institute.* The variant forms can be described as text occurrences conceptually well related with the correct form or the standardized form. Table 1 shows just a few variant forms of a single address – a university department – as stored in the citation databases of Institute for Information Science (ISI) through the Web of Science (THE THOMSON CORPORATION, 2005).

Table 1. Some variants of a single institutional address downloaded from the ISI Web of Science

Univ Granada, Comp Sci & Artificial Intelligence Dept, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Comp & IA, ETSI Informat, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Comp & IA, Granada 18071, Spain.
Univ Granada, Dept Ciencias Comp IA ETSII Informat, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac & IA, ETS Ingn Informat, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac & IA, ETSI Informat, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac & IA, Granada 18071, Spain.
Univ Granada, Dept Ciencias Computac & Inteligencia Artificial, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac & Inteligencia Artificial, ETSI Informat, Granada 18071, Spain.
Univ Granada, Dept Ciencias Computac & Inteligencia Artificial, Granada 18071, Spain.
Univ Granada, Dept Ciencias Computac & Inteligncia Artificial, Granada, Spain.
Univ Granada, Dept Ciencias Computac & JA ETSII, Granada 18071, Spain.
Univ Granada, Dept Ciencias Computac & LA, Granada 18071, Spain.
Univ Granada, Dept Ciencias Computac e IA, ETS Ingn Informat, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac EIA, ETS Ingn Informat, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac IA, E-18071 Granada, Spain.
Univ Granada, Dept Ciencias Computac, EIA ETSI Informat, E-18071 Granada, Spain.
Univ Granada, Dept Comp Sci & AI, E-18071 Granada, Spain.
Univ Granada, Dept Comp Sci & AI, Granada 18071, Spain.
Univ Granada, Dept Comp Sci & Artif Intelligence, E-18071 Granada, Spain.
Univ Granada, Dept Comp Sci & Artificial Intelligence, E-18071 Granada, Spain.
Univ Granada, Dept Comp Sci & Artificial Intelligence, ETSI Informat, E-18071 Granada, Spain.
Univ Granada, Dept Comp Sci & Artificial Intelligence, Granada, Spain.
Univ Granada, Dept Comp Sci, E-18071 Granada, Spain.
Univ Granada, Dept Comp Sci, Granada, Spain.
Univ Granada, Dipartimento Ciencias Computac & Inteligencia Art, E-18071 Granada, Spain.
Univ Granada, Dpto Ciencias Computac & Inteligencia Artif, ETSI Informat, E-18071 Granada, Spain.
Univ Granada, Escuela Tecn Super Ingn Informat, Dept Ciencias Computac & IA, E-18071 Granada, Spain.
Univ Granada, ETS Ingn Informat, Dept Ciencias Comput & IA, E-18071 Granada, Spain.

These inconsistencies result from:

- *Non-acceptable variations*, including non-valid addresses, or incorrect variant forms. The reason behind such variations would essentially be errors, misspelled words and inaccurate translations of foreign terms.

* The problem with departmental addresses is not only the lack of standardization, but also that departmental addresses may be incomplete, and departmental structures are not stable over time. The impact or usefulness of departments (as opposed to main institutions) in bibliometric analyses is limited by the fact that this information is sometimes simply omitted.

- *Acceptable variations*, which would be valid addresses, or correct variant forms. Here the most frequent causes are permuted word order or distinct syntactic formats of the same name, the splitting of words, acronyms, full vs. abbreviated address, transliteration differences, differences in US versus UK spelling, the inclusion or exclusion of postal codes, state and country names, inclusion or not of the main organization or research group names. Overall, these addresses are interchangeable in specific contexts without leading to a change in meaning.

The standardization of these institutional affiliation data looms as a very complicated and time-consuming operation. Whereas on the one hand the city information can be rather easily standardized by eliminating the postal codes, the main organization may have a great number of variants, not to mention the departments in that organization. To solve these problems we can use approximate string matching (HALL & DOWLING, 1980), which entails essentially two procedures: (*i*) based on similarity relations between the non-valid variants and the correct one(s) using measures of similarity; and (*ii*) based on equivalence relations between the valid variants and canonical forms, which requires the computation of the equivalence relation.

Our main purpose is to present a procedure that would automatically standardize the university department addresses in databases on the basis of equivalence relations. We focus on the group of valid variants that seek to map multiple variations into a single class, defined as the canonical form, in turn divisible into two separate approaches: clustering techniques vs. finite-state techniques.

Equivalence classes based on clustering techniques. In this first group, we find studies such as that of FRENCH et al. (2000), where it is attempted to assign to each affiliation address a canonical form through the application of clustering algorithms. Basically, the stages called for are: cleaning, sorting, clustering, checking and updating. The addresses must first be put through a lexical cleanup process, where abbreviations and acronyms are expanded, accents are removed, or the shift string is transformed to lower-case. Then the addresses are sorted, in descending order, by frequency of appearance. The most frequent address is selected as the canonical form and is compared with the rest of the addresses, using a measure of similarity; this process is repeated successively until all the strings have been clustered. The resulting clusters are verified, with all the possible errors localized and corrected. Finally, the corporate source data are updated and the addresses of a cluster are replaced by its standardized form.

Equivalence classes based on finite-state techniques. In this study, we introduce a procedure based on approximate parametrized matching through transducers. We shall consider the valid addresses as patterns or frozen expressions that will be identified by

means of finite-state methods, entailing the application of parsing techniques. To begin, the problem of the unification of institutional data must be formalized in terms of units of analysis and a new data structure, binary matrices, must be defined. Then, the standardization process will be carried out by applying master graphs, transformed into transducers. The ultimate objective of this research is to present a Natural Language Processing (NLP) oriented method for the standardization of corporate source data by semi-automatic means, requiring less human interaction, and easily adapted to bibliometric applications.

Approach to standardization processes applying transducers

Transducers have been used for multiple tasks in computer sciences, in NLP and Information Retrieval (IR). In the last two decades mathematical procedures have allowed for very significant advances, with demonstrated efficacy in pattern recognizing, tokenization, lexical analysis, parsing, conflation algorithms and the standardization of personal names (KARTTUNEN et al., 1992; SILBERZTEIN, 1993; KAPLAN & KAY, 1994; ROCHE & SCHABES, 1995; ABNEY, 1996; MOHRI, 1996; AIT-MOKHTAR & CHANOD, 1997; JACQUEMIN & TZOUKERMANN, 1999; GALVEZ & MOYA-ANEGÓN, in press).

Transducers are an extension of finite automata (HOPCROFT & ULLMAN, 1979), or mathematical models of a system with input and output, and can be defined as a finite set of states and a set of transitions from one state to another. Transducers define relations between languages. To compute the relations, a transducer has transitions labeled with two symbols from two alphabets: input and output. Formally, a finite-state transducer (FST) is characterized as a 5-tuple, $T = (Q, \Sigma, q_0, F, \delta)$, where Q is a finite set of states, Σ is the input and output alphabet (and E is the empty string), q_0 is the initial state, F is the set of final states, and δ is the set of transitions (ROCHE & SCHABES, 1995). The transducers can be represented as directed graphs, whose vertices denote states, while the transitions form the edges, or arcs, with arrows pointing from the initial state to the final state.

Using a graphic interface, we drew finite-state graphs that would represent the possible structures underlying variants of a departmental address, and could produce as output the format selected as the standardized form of this department (see Figure 1). In this case, we prioritize a format used by the ISI, which brings some order in addresses – although we could have opted for any other particular one – with four components: <University, ISI-Standardized Abbreviations, City, Country>.

The finite-state graph, in this case built with the interface FSGraph (SILBERZTEIN, 1993; 2000), is compiled into a FST; this application allows for the graph's

transformation into a table or transition matrix (see Table 2) where the following components are specified:

- Number of states, $Q = 23$.
- Number of alphabet symbols, or vocabulary, $\Sigma = 22$, where the symbol <*E*> represents the empty string.
- Initial state, $q_0 = 0$.
- Final state, $F = 1$.
- Number of transitions between states, $\delta = 27$, where each transition is defined by a 3-tuple: current state, symbol, outgoing state.

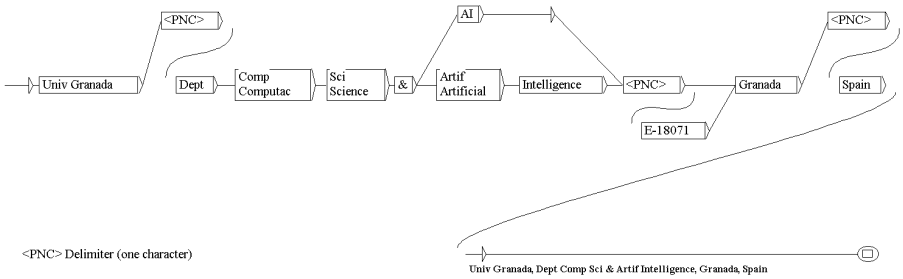

Figure 1. Finite-state graph developed for grouping the address variants into a standardized format

The FST obtained recognizes 24 variant formats of the address 'University of Granada, Department of Computer Science and Artificial Intelligence, Granada, Spain'. These variants would belong to the same equivalence class – characterized as a representative member of the class defined by the canonical form 'Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain' (see Table 3). Nevertheless, this procedure would be useless, as it would need the hand-drawn representation of each equivalence class and the thousands of variants that these sorts of sequences might have. The next sections of the paper will describe a semi-automatic approach that would allow us to recognize and unify the valid variants into standard forms.

Table 2. Result of the transformation of the finite-state graph into a transition table

// Characteristics of the FST
#define NUMBER_OF_STATES 23 // states are numbered from 1 to NUMBER_OF_STATES
#define NUMBER_OF_SYMBOLS 22 // symbols are numbered from 0 to NUMBER_OF_SYMBOLS-1
#define NUMBER_OF_TRANSITIONS 27
// FST Alphabet/Vocabulary
static const char *symbols[NUMBER_OF_SYMBOLS]={"E","AI","Science","Artif","Sci","<E>/Univ Granada, Comp Sci & Artif Intelligence, Granada, Spain",",","&","-","1","8","0","7","Computac","Comp","Dept","Univ","Intelligence","<E>","Artificial","Granada","Spain"};
// FST terminal states: 0=non terminal; 1=terminal
static const char terminal_state[NUMBER_OF_STATES+1]={'\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\1'};
// FST transitions: each transition is a 3-tuple (current state,symbol,outgoing state)
static const int transitions[NUMBER_OF_TRANSITIONS*3]={
1,16,2,
2,20,3,
3,6,4,
4,15,5,
5,13,6,5,14,6,
6,2,7,6,4,7,
7,7,8,
8,1,9,8,3,10,8,19,10,
9,6,11,
10,17,12,
11,0,13,11,20,14,
12,6,11,
13,8,15,
14,6,21,
15,9,16,
16,10,17,
17,11,18,
18,12,19,
19,9,20,
20,20,14,
21,21,22,
22,5,23,
};

Table 3. Standardized forms generated by the finite-state graph

Univ Granada, Dept Comp Science & AI, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Comp Sci & AI, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Science & AI, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Sci & AI, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Sci & Artificial Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Sci & Artif Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Science & Artificial Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Science & Artif Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Comp Sci & Artificial Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Comp Science & Artificial Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Comp Science & Artif Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Computac Sci & AI, E - 1 8 0 7 1 Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
/.../	

Methodology

To solve the above issue, we consider the problem of standardization by adopting the concept of parametrized matching (p-matching)* and the application of finite-state methods. Assuming that transducers can be used to represent equivalence relations, the problem of standardization can be viewed as a equivalence relation that maps institutional variants to standardized structures. We therefore formalize this problem in terms of p-matching, through p-graphs defined as finite-state graphs compiled in transducers, whose alphabet of input and output contains parameters with values that depend on entries given in an *ad hoc* matrix. The parameters refer to the contents of the table by means of variables, in such a way that the p-matching is done by surrogating variables. Using a linguistic development environment based on finite state technology (SILBERZTEIN, 2000), we try to determine whether two institutional addresses that are theoretically correct can be made into a p-match, establishing an equivalence relation through substitutions for variables.

The methodology presented here involves three stages: (1) a thorough examination of the components of departmental names in terms of independent units of analysis (UA); (2) the representation of these components in a binary matrix; and (3) the identification and posterior standardization of the constructions by means of a Parametrized Finite-State Graph (P-FSG), compiled in a transducer.

* This notion was introduced by BAKER (1993) for applications that arise in software tools for analysing source codes. A parameterized (or parametrized) string matching is a string over the union of two alphabets (an alphabet Σ of constant symbols and an alphabet Π of parameter symbols); then "two p-strings are a parameterized match, or p-match, if one p-string can be transformed into the other by applying a one-to-one function that renames the parameter symbols" (BAKER, 1996, p. 28). The identity of the constant symbols and a permutation of the parameter symbols, according to the initial proposal by BAKER (1996), were calculated using parameterized suffix trees, or p-suffix trees, defined as a tree structure where each node represents one character and the root represents the null string.

For model constructions, we take a sample of address structures from the Science Citation Index Expanded (SCI-E) database. The collection need not be very large, as there are only so many legitimate structural forms of addresses, and so after a certain point, the larger the sample, the lesser the variations. The dataset was gathered from all papers with the term 'University of Granada' and the word 'Department' in the research address field. The choice of SCI-E databases is justified by the fact that they contain the list of all addresses indicated in the publications.

We implemented this prototype system for the standardization of a very specific problem, as is the unification of address data, but the system can be expanded to other possible applications. By means of linguistic development environments based on finite-state technology (SILBERZTEIN, 2000; PAUMIER, 2003), users can add their own resources and use these tools for applications such as the unification of variants of proper names, journal titles and other corporate data.

Units of analysis (UA)

In the corporate source field of ISI databases, the institutional references usually contain: the name of the overall organization, such as universities; sub-organizations or divisions, such as faculties or institutes; and subdivisions, such as departments or research groups (DE BRUIN & MOED, 1993). However, this supposedly hierarchical order (<OG>Organization/<SG>Sub-organizations) does not always prevail, and we may encounter cases where the overall organization is omitted and a subdivision is given directly (e.g., 'Dept Biochem & Mol Biol, Granada, Spain'), or cases where a single subdivision appears linked to two different divisions (e.g., 'Univ Granada, *Fac Med*, Dept Biochem & Mol Biol, Granada, Spain' and 'Univ Granada, *Fac Pharm,* Dept Biochem & Mol Biol, Granada, Spain'). This type of situation, together with the problems stemming from the permuted order of institutional parts, has led us to adopt the notions of independent units of analysis (UA) and aggregation levels (AL) to focus the affiliation-related technical problems.

In bibliometric research, units of analysis are well-known notions and are defined as the objects of study described by variables about which inferences are made (MOHR, 1990; MCGRATH, 1996); and the selection of these units depends on the focus in question (e.g., the scientific publications are the variables assigned to institutions, departments or countries as the units of analysis, through the corporate addresses of their authors). In turn, depending on the aggregated and deaggregated levels of the unit of analysis on which one chooses to focus, the bibliometric research can be performed at different aggregation levels, classified by VAN RAAN (2003) as: (1) the macro-level of analysis, such as entire countries or regions; (2) the meso-level of analysis, examining larger institutions such as universities, or their major parts, like schools,

faculties or institutes; and (3) the micro-level of analysis, investigating departments and research groups within universities and institutes; on this level, the necessary information is available only within the university or institute itself and must always be gathered separately.

Given that our purpose is to develop a model for the unification of corporate source data that would have a possible bibliometric application in the future, the present study calls for come consideration of problems regarding the unification of the names of departments in terms of independent units of analysis at the lower aggregation level. Our approach is based on the following sequential order of the corporate infrastructure data in ISI databases, tagged as:

- Meso-level: *UA*1 (University), *UA*2 (Faculty/Institute).
- Micro-level: *UA*3 (Department), *UA*4 (Research Group).
- Macro-level: *UA*5 (City), *UA*6 (Country).

Thus, the information about the departments would be 'broken', and at the same time integrated, into other units in an independent manner, as university or country, without establishing any hierarchy among them (e.g., '<*UA*1>Univ Granada, <*UA*2>Fac Med, <*UA*3>Dept Biochem & Mol Biol, <*UA*5>Granada, <*UA*6>Spain'). One important advantage of this proposal is that the same scheme could be used as a model guiding the unification of other units, such as the name of the same university, institute, hospital, or city at a higher aggregation level, or the unification of research group, team, or smaller communities of researchers and programs at a lower aggregation level.

Binary matrix

A binary matrix or lexicon-grammar matrix is a theoretical model, lending itself particularly well to NLP, which describes languages in a systematic way. This formalism was introduced by GROSS (1975; 1997) with the original goal of a linguistic description having sufficient scientific rigour of given expressions. The association between syntax and lexicology led to the birth of this model, in which the linguistic data studied are presented within a binary matrix or table. Adopting Gross's model of description, the departmental data considered in the frame of lexicon-grammar are codified in a binary matrix, using a spreadsheet application. Normally, each line should represent a departmental entry – however, in our case, two lines were needed for each departmental entry, because of the use of two different languages, Spanish and English.

The columns of the matrix correspond to an address part, or a property. We encoded the data of the given departmental entry into a binary matrix with the following format:

- The first line of the column contains:

a) The name of a part of the unit of analysis studied, in this case, the name of the department, *UA*3 (Department), such as *UA* 3-2 (Part 2 of the name of the department), *UA*3-3 (Part 3 of the name of the department), *UA*3-4 (Part 4 of the name of the department), etc. Moreover, as on a micro-level the information is offered within the university, *UA*1 (University), we decided to include as well the name of the parts of that entity, such as *UA*1-1 (Part 1 of University), *UA*1-2 (Part 2 of University). Additionally, we include the names of the city, *UA*5 (City), and country, *UA*6 (Country). Then we add the standardized abbreviations used in ISI databases, *ISI – SA* (ISI-Standardized Abbreviations), for reasons we explain below.

b) The name of a property of the unit of analysis studied, in this case the department address, such as *UA*3-1 (Abbreviations of the term 'Department' and the use of this term in foreign languages). Also included are properties of other units of analysis, such as *UA*5-1 (Zip/Postal Code), *UA*5-2 (Province/State). We furthermore decided to include as properties those units of analysis that would not be the object of study but that might appear in conjunction with the departmental address, such as *UA*2 (Faculty/Institute) or *UA*4 (Research Group).

- The column-line intersections of a text cell, corresponding to a part or zone of the name of department, are filled with constants (lexical elements) or else with an empty string represented by the symbol <*E*>.
- The column-line intersections of a property cell are filled with a symbol (+)if the current entry validates the institutional property, or with the symbol (–)if it does not.
- Each departmental entry is described in a single line, and the contents of the text and property cells must be homogenous and cannot be empty.

This matrix affords a uniform representation of the features corresponding to the possible departmental structures considered. In a synthesized form, the theoretical conception that underlies the binary matrix would be to consider the components of a system of formal rules that explicitly assign a description of specific constructions. In the table, there would be no dissociation between the names of parts and properties (syntactic components such as prepositions) and the constants (lexical components such as 'Univ' or 'Granada'). The advantages of this data structure are that it is clearly oriented to computational processing.

In view of this notion, we consider that the possible structures of departmental constructions are frozen expressions, and the grammars that characterize these expressions will be built indirectly, by means of binary matrices. For the description of such structures, we built a 144 x 28 matrix, an extract of which is given in Table 4 (in

the Appendix). This formalism is a simple representation and cannot be used as a mechanism of identification of such structures, but only to classify them. Hence, it is necessary to build a transducer that is able to identify and unify these constructions by means of a new syntactic parser, based on the association of a finite-state graph with the data in the binary matrix (ROCHE, 1993; 1996).

Parametrized finite-state graph (P-FSG)

In order that the departmental information encoded in the matrix can be computationally processed, it must first be transformed into a transducer. This process essentially involves associating it to a finite-state graph, or master graph. The master graph will be built manually, using a graphic interface with parameters whose values correspond to a feature of the matrix. With this idea in mind, we may consider creating a P-FSG that would check all its features: constants (lexical elements) and property codings (+ or –).

This master graph would represent the set of all possible forms and refer to the content of the matrix by means of variables (where the variable *@A* refers to the content of the cell found at the intersection of a specific line and the column *A* of the matrix, the variable *@B* refers to the content of the cell found at the intersection of the line and column *B*, etc.). Thus, for each line of the matrix, or departmental entry, there would be a copy in the master graph that would automatically adjust to the contents of the matrix, through the variables, in the following fashion:

- Replacing the variable with the content of the column-line intersection of a text cell, that is, either with a constant or an empty string (<*E*>).
- Maintaining the path or transition if the variable refers to the line-column intersection of a property cell filled with the symbol (+).
- Removing the path or transition if the variable refers to the line-column intersection of a property cell filled with the symbol (–).

This analysis then combines a binary matrix, which formalizes all the constructions to a departmental entry, and a finite-state graph with parameters, according to the departmental entries stored in this matrix. Using this procedure we could identify all the possible valid variants of the departmental addresses, though not standardize them. In order, then, to unify this type of construction, we propose the introduction of a parametrized transducer, whose alphabet of input and output is made up of variables. Standardization would thus involve:

- Adding variables in the output of the parametrized graph, associated with standardized entries in the matrix.

- Replacing the values of the variables recognized in the input with values of the variables in the output, corresponding to the values held to be the standardized forms, representative of the equivalence class.

In turn, this parametrized graph can be represented in the form of an Enhanced Finite-State Transducer (EFST), defined as a transducer that contains, besides, internal variables used during the parsing to order or classify the parts of the sequences recognized. In this way, the input sequences that match are indexed simultaneously with the corresponding output. That is, in EFST outputs and inputs are synchronized by means of internal variables to store parts of the matching input sequence: the internal variables are introduced as tagged parentheses around the corresponding sequences and are identified by the symbol (*$*). The use of EFST variables permits establishing the order of the recognized sequences, and making any necessary permutations or insertions. It also allows us to intentionally modify the conditions of this synchronization in order to obtain the correct matching forms. The variables that we propose for the classification of the departmental sequences have been targeted as:

- *$UA*1 (University);
- *$UA*2-*UA*4 (Faculty/Institute – Research Group);
- *$UA*3 (Department);
- *$UA*5 (City);
- *$UA*6 (Country).

Finally, the parametrized graph is compiled in a transducer in charge of determining whether two expressions can be made equivalent via substitutions for variables. As a result, all operations defined for the graph are also defined for instances of the matrix. In Figure 2, we show an extract of P-FSG in charge of recognizing departmental address variants that matches them with those elements of the matrix that are to form part of the canonical sequence, in this particular case defined by ISI standards for address abbreviations.

Evaluation

This procedure was tested on a data collection downloaded from the SCI-E through the Web of Science (THE THOMSON CORPORATION, 2005). The dataset covers the period 2003-2004, and a total of 1507 records were obtained with the terms 'Univ Granada' and 'Dept not Hosp' in the address field (AD). The data of the study were collected together in relation to the overall organization concerned, 'University of Granada', because our intention was to have a sampling for later evaluation. The set of references obtained was imported to a bibliographic management system, ProCite database (version 5.0), in order to automatically generate a list of variants that could be

quantified in the evaluation and eliminate the duplicate addresses, leading to a reduction to 719 different addresses. Before attempting analysis, the list was put through a series of transformations so that it could be processed in the text-file format and segmented into sentences. The next step was to apply the parametrized transducer to the occurrences of the selected addresses. An extract of the data obtained is given in Table 5.

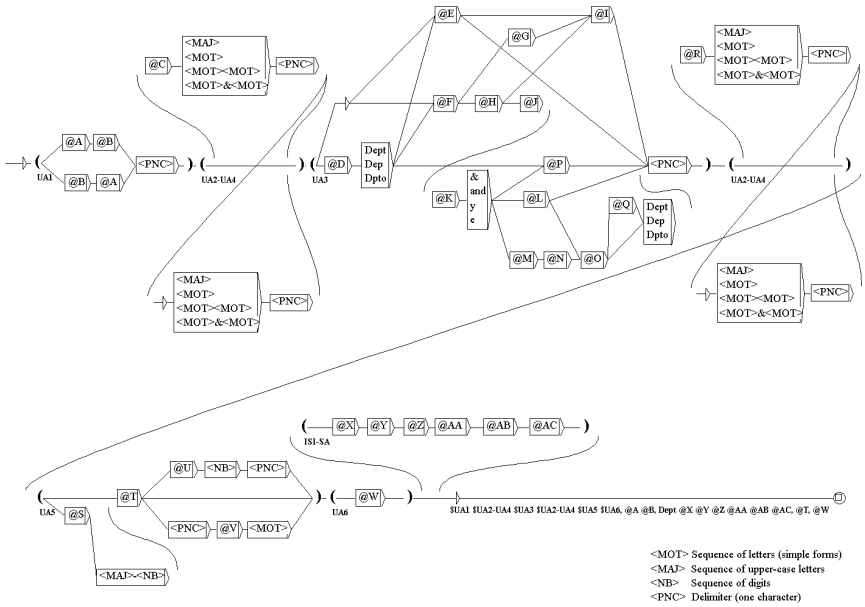

Figure 2. Simplified version of the P-FSG built for the unification of departmental addresses

Table 5. Excerpt of data obtained with the application of the P-FSG in a selection of departmental addresses

Univ Granada, Dep Arquitectura & Tecnol Computadores, E-18071 Granada, Spain =>	**Univ Granada, Dept Architecture & Comp Technol, Granada, Spain**
Univ Granada, Dept Anal Matemat, E-18071 Granada, Spain =>	**Univ Granada, Dept Math Anal, Granada, Spain**
Univ Granada, Dept Anat & Patol, Fac Med, Granada, Spain =>	**Univ Granada, Dept Anat & Pathol, Granada, Spain**
Univ Granada, Dept Biol Vegetal, Fac Ciencias, E-18071 Granada, Spain =>	**Univ Granada, Dept Plant Biol, Granada, Spain**
Univ Granada, Dept Comp Sci & AI, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Comp Sci & Artificial Intelligence, Granada, Spain =>	**Univ Granada, Dept Comp Sci & Artif Intelligence, Granada, Spain**
Univ Granada, Dept Bioquim & Biol Mol, Granada, Spain =>	**Univ Granada, Dept Biochem & Mol Biol, Granada, Spain**
Univ Granada, Dept Electromagnetismo, Fac Ciencias, E-18071 Granada, Spain =>	**Univ Granada, Dept Electromagnetism & Mat Phys, Granada, Spain**
Univ Granada, Dept Estadist & Invest Operat, E-18071 Granada, Spain =>	**Univ Granada, Dept Stat & Operat Res, Granada, Spain**
Univ Granada, Dept Estrat & Paleontol, Granada 18002, Spain =>	**Univ Granada, Dept Strat & Paleontol, Granada, Spain**
Univ Granada, Fac Ciencias, Dept Fis Aplicada, Grp Fis Fluidos & Biocoloides, Granada 18071, Spain =>	**Univ Granada, Dept Appl Phys, Granada, Spain**
Univ Granada, Fac Ciencias, Dept Ingn Quim, E-18071 Granada, Spain =>	**Univ Granada, Dept Inorgan Chem, Granada, Spain**
Univ Granada, Fac Farm, Dept Nutr & Bromatol, E-18012 Granada, Spain =>	**Univ Granada, Dept Nutr & Food Sci, Granada, Spain**
Univ Granada, Fac Med, Dept Bioquim & Biol Mol, Granada, Spain =>	**Univ Granada, Dept Biochem & Mol Biol, Granada, Spain**
Univ Granada, Fac Med, Dept Fisiol, E-18012 Granada, Spain =>	**Univ Granada, Dept Physiol, Granada, Spain**
Univ Granada, Fac Pharm, Dept Quim Inorgan, E-18071 Granada, Spain =>	**Univ Granada, Dept Inorgan Chem, Granada, Spain**
Univ Granada, Fac Sci, Dept Plant Biol, Granada 18071, Spain =>	**Univ Granada, Dept Plant Biol, Granada, Spain**
Univ Granada, Unidad Inmunol, Dept Bioquim & Biol Mol, Fac Med, Granada 18012, Spain =>	**Univ Granada, Dept Biochem & Mol Biol, Granada, Spain**
/.../	

For the evaluation of effectiveness, we used an adaptation of precision and recall measures based on accuracy and coverage, not actual retrieval. In this context, recall would normally indicate the proportion of terms that are standardized with respect to a set of sequences of evaluation, yet we modify its definition slightly to stand for 'correct variants standardized over total possible variants susceptible of unification'. Precision is in turn understood as the ratio of correct variants standardized from among the total variants standardized by the finite-state graph. The two measures were calculated through the following equations:

$$\text{Recall (R)} = \frac{\text{Number of Correct Variants Standardized}}{\text{Total Number of Possible Variants}}$$

$$\text{Precision (P)} = \frac{\text{Number of Correct Variants Standardized}}{\text{Total Number of Variants Standardized}}$$

We then applied a measure of performance that takes into account both recall and precision: the F-score (VAN RIJSBERGEN, 1979) defined as the harmonic mean of recall and precision, as compared to the arithmetic mean, which exhibits the desirable properties of being highest when both recall and precision are high. The variable β weights the relative importance of both (β = 1 means that recall and precision are equally weighted; whereas $\beta > 1$ means more weight for recall, and $\beta < 1$ means more weight for precision). We established a value of β = 1 intended for to treat the two equally. Calculation is as follows:

$$F_\beta = \frac{(\beta^2 + 1)RP}{\beta^2 R + P}$$

In order to arrive at these figures, we needed to have from the following data of frequency: (a) number of correct variants standardized (the output of P-FSG as compared with the words that had been successfully standardized, removing under-standardized errors and over-standardized errors); (b) total number of possible variants (total variants that should have been grouped to a standardized form), these data were obtained manually; (c) total number of variants standardized (total number of unique addresses minus number of variants not standardized). The percentage of under-standardization and over-standardization errors* could then be calculated as follows:

$$\text{Under-standardization Errors} = \frac{\text{Number of Variants not Standardized}}{\text{Total Number of Possible Variants}}$$

* The under-standardization errors occur when address names that refer to the same variants are not reduced to the same unified format; and over-standardization errors occur when address names are standardized incorrectly because they are not actual variants.

$$\text{Over-standardization Errors} = \frac{\text{Number of Nonvariants Standardized}}{\text{Total Number of Variants Standardized}}$$

Results and discussion

An analysis of the results of standardizing university department addresses through the finite-state transducers (see Table 6) shows that the variants are identified with a recall of $R = 0.87$. The non-standardization rate is 0.07, below the baseline $F_1 = 0.94$. This lack of recall or under-standardization occurs because the p-graph can only detect values specified in the matrices. The percentage of under-standardized addresses was 0.12, these owing to non-valid variants, including:

- University department names with spelling errors (e.g., 'Dept Algegra' for 'Dept Algebra'). The misspelling percentage was found to be 0.23.
- Non-authentic department addresses, usually stemming from different addresses that appear united, as if pertaining to a single institutional affiliation (e.g., 'Dept Med & Org Chem' for 'Dept Med' and 'Dept Org Chem'). The non-legitimate address percentage was 0.29.
- Department names that do not coincide with the values stored in the matrix for any number of phenomena (incorrect denominations such as 'Dept Comp Technol & Comp Architecture' for 'Dept Architecture & Comp Technol', scattering of organizations or the inappropriate attachment of a department to an overall organization when in reality it belongs to a suborganization, for instance, 'Univ Granada, Dept Immunol, Granada, Spain' for 'Univ Granada, *Hosp Virgen Nieves,* Dept Immunol, Granada, Spain') and departmental addresses that do not actually exist as such, often due to inaccurate translations of foreign terms, or else owing to errors in the primary literature itself (e.g., 'Dept Lib Sci Studies' for 'Dept Lib & Informat Sci'). This mismatching address percentage was found to be 0.48.

A variety of approaches and procedures might be used to remedy such inaccuracies. The first type of error could be resolved with spelling correction techniques and approximate matching methods based on similarity relations (between valid and non-valid variants). The second type might be avoided altogether through a pre-processing stage: once the data is downloaded, offline correction is performed until it is clear that reference is made to separate institutions. This leaves us with the more imposing third type of failures, those stemming from inconsistencies produced by departmental name-changes or erroneous combinations of address components. Such a situation implies a complex problem of ambiguity; though a single institution may be cited in a number of ways, not all can be accepted as valid variants. Therefore, disambiguation calls for

methods that distinguish the contexts in which names appear, such as using the vector space model to resolve ambiguities, or co-occurrence analysis and clustering techniques based on contextual similarities. Here we would need to resort to semantic knowledge plus world knowledge. At this point in time, however, our only objective is to apply the prototype to identify the equivalent address valid variants and formats that indeed refer to the same name, and the problems of ambiguity must therefore be relegated beyond the scope of the present contribution.

Table 6. Recall, precision and F-score measures

Possible Variants	719
Variants Standardized	631
Non-valid Variants	88
Correct Variants Standardized	631
Recall	**0.87**
Precision	**1.00**
F_1	**0.94**

According to our evaluation, the address variants are identified with a remarkably high precision of $P = 1$, well above the baseline of $F_1 = 0.94$. This very high precision index can be explained by the *ad hoc* nature of the assignment at hand: all the possible address formats, foreign names and abbreviations to be identified by the P-FSG were previously and exhaustively specified in the binary matrix. No instance of over-standardization was seen, because P-FSG does not process non-valid variants as if they were valid ones.

A further explanation of computational efficacy resides in the fact that, when applying P-FSG to address sequences, we can choose to index 'shortest matches', 'longest matches', or 'all matches'. If address names of varying lengths are identified, with sequences occupying the same initial position, the application itself allows us to select a prioritized mode of recognition of the address name variants. We specified only 'longest matches' to be taken into consideration by the system, obtaining 631 acknowledged sequences. If, for example, 'all matches' had taken priority over the longer ones, the total number of variants recognized would have been 744, entailing a negative impact on precision.

In contrast, efficiency in precision is related to a basic drawback of this approach: no one knows exactly what constructions might appear in a collection, and yet we must mention *a priori* the type of address formats to be processed before beginning to

develop the model. The options are therefore to either devise a list of the constructions that we propose to process, or else limit the task to real examples taken from the domain of application, with the understanding that the collection need not be large, because the proportion of formats diminishes as the sample increases.

Conclusions

Advanced bibliometric methods oriented particularly at the level of research groups, university departments and institutes, identified in the address field, are an indispensable element in research evaluation procedures (VAN RAAN, 1999). Many variations occur with respect to the names of organizations, in the analysis of addresses in scientific publications, and this phenomenon has serious consequences for the availability of information (DE BRUIN & MOED, 1990), as a result the need to standardize corporate source data will be crucial in performance measurement (HOOD & WILSON, 2003). While databases provide the raw resources for bibliometric studies, a number of obstacles must be overcome to ensure quality in bibliometric searches, and one major difficulty lies in the lack of consistency among data at the micro-level. Our proposal revolves around a procedure based on parametrized matching and finite-state techniques that could unify corporate source data. On the basis of the experiments performed and described here, three significant conclusions can be drawn. First, the results of this procedure in terms of recall point to a problem of under-standardization, accentuated by strings that do not coincide with the values stored in the matrix. Second, the precision of the P-FSG in the analysis and recognition of address variants is complete, producing no over-standardization errors. This is because recognition is strongly guided by the parameters contained in the binary matrices and because we gave priority to longest sequence matching. Third, disambiguation of the contexts in which institutional addresses appear remains a point of contention that requires complementary solutions involving similarity measures, co-occurrence analysis and clustering techniques; if not, the only solution to these problems is to perform offline manual correction.

The model presented here describes the corporate source data at micro-level in a systematic form. It attempts to solve the problem of departmental address structures that may appear in varied formats, dealt with as a combination of variables in which any constants could be inserted. Moreover, it may resolve inconsistencies produced by abbreviations or the translation of foreign names. In terms of manpower savings, although substantial resources are necessary during the first stage of the process, this is compensated by the subsequent benefits because the tools can be utilized again. The main contribution of this approach is that it provides for a uniform description, representation and identification of this type of sequences. Standardization of variant formats through P-FSG stands as a potential solution to the sorts of problems that arise

when databases are used for bibliometric purposes. Human interaction will always be necessary to some degree, in the pre-processing stage and in the creation of *ad hoc* binary matrices; yet with this procedure, much time and effort might be saved, and precision gained, in the extraction of information from databases for quantitative studies.

References

ABNEY, S. (1996), Partial parsing via finite-state cascades, *Natural Language Engineering*, 2 : 337–344.

AIT-MOKHTAR, S., CHANOD, J. (1997), Incremental finite state parsing. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. ACL, pp. 72–79.

ANDERSON, J., COLLINS, P. M. D., IRVINE, J., ISARD, P. A., MARTIN, B. R., NARIN, B. R., STEVENS, K. (1988), On-line approaches to measuring national scientific output: A cautionary tale, *Science and Public Policy*, 15 : 153–161.

BAKER, B. S. (1996), Parameterized pattern matching: Algorithms and applications, *Journal of Computing and System Sciences*, 52 : 28–42.

BAKER, B. S. (1993), A theory of parameterized pattern matching: Algorithms and applications (extended abstract). In: *Proceedings of the 25th Annual Symposium on Theory of Computing*. ACM Press, pp. 71–80.

BOURKE, P., BUTLER, L. (1996), Standards issues in a national bibliometric database: The Australian case, *Scientometrics*, 35 : 199–207.

BOURKE, P., BUTLER, L. (1998), Institutions and the map of science: Matching university departments and fields of research, *Research Policy*, 26 : 711–718.

BRAUN, T., BROCKEN, M., GLÄNZEL, W., RINIA, E., SCHUBERT, A. (1995), "Hyphenation" of databases in building scientometric indicadors: Physics Briefs – SCI based indicators of 13 European countries, 1980–1989, *Scientometrics* 33 : 131–148.

CARPENTER, M. P., GIBB, F., HARRIS, J., IRVINE, J., NARIN, F. (1988), Bibliometric profiles for British academic institutions: An experiment to develop research output indicators, *Scientometrics*, 14 : 213–234.

CRONIN, B., SNYDER, H. W. (1997), Comparative citation ranking of authors in monographic and journal literature: A study of sociology, *Journal of Documentation*, 53 : 263–273.

DE BRUIN, R. E., MOED, H. F. (1993), Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications, *Scientometrics*, 26 : 65–80.

DE BRUIN, R. E., MOED, H. F. (1990), The unification of addresses in scientific publications. In: L. EGGHE, R. ROUSSEAU (Eds), *Informetrics 1989/90*. Elsevier Science Publishers, Amsterdam, pp. 65–78.

FRENCH, J. C., POWELL, A. L., SCHULMAN, E. (2000), Using clustering strategies for creating authority files, *Journal of the American Society for Information Science and Technology*, 51 : 774–786.

GALVEZ, C., MOYA-ANEGÓN, F. (2006), Approximate personal name-matching through finite-state graphs. *Journal of the American Society for Information Science and Technology*, in press.

GARFIELD, E. (1979), *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*, John Wiley, New York.

GARFIELD, E. (1983a), Idiosyncrasies and errors, or the terrible things journals do to us, *Current Contents*, 2 : 5–11.

GARFIELD, E. (1983b), Quality control at ISI, *Current Contents*, 19 : 5–12.

GILES, C. L., BOLLACKER, K., LAWRENCE, S. (1998), CiteSeer: An automatic citation indexing system. In: I. WITTEN, R. AKSCYN, F. M. SHIPMAN III (Eds), *Digital libraries 98 - The Third ACM Conference on Digital Libraries*. ACM Press, pp. 89–98.

GROSS, M. (1975), *Méthodes en Syntaxe*, Hermann, Paris.

GROSS, M. (1997), The construction of local grammars. In: E. ROCHE, Y. SCHABES (Eds), *Finite-State Language Processing*. MIT Press, pp. 329–352.

HALL, P. A. V., DOWLING, G. R. (1980), Approximate string matching, *Computing Surveys*, 12 (4) : 381–402.

HERBERTZ, H., MÜLLER-HILL, B. (1995), Quality and efficiency of basic research in molecular biology: A bibliometric analysis of thirteen excellent research institutes, *Research Policy*, 24 : 959–979.

HOOD, W. W., WILSON, C. S. (2003), Informetric studies using databases: Opportunities and challenges. *Scientometrics*, 58 : 587–608.

JACQUEMIN, C., TZOUKERMANN, E. (1999), NLP for term variant extraction: Synergy between morphology, lexicon, and syntax. In: T. STRZALKOWSKI (Ed.), *Natural Language Information Retrieval*. Kluwer Academic Publishers, Dordrecht, pp. 25–74.

LEYDESDORFF, L. (1988), Problems with the 'measurement' of national scientific performance, *Science and Public Policy*, 15 : 149–152.

MÄHLCK, P., PERSSON, O. (2000), Socio-bibliometric mapping of intra-departmental networks, *Scientometrics*, 49 : 81–91.

MCGRATH, W. (1996), The unit of analysis (object of study) in biblometrics and scientometrics, *Scientometrics*, 32 : 257–264.

MELIN, G., PERSSON, O. (1996), Studying research collaboration using co-authorships, *Scientometrics*, 36 : 363–377.

MOHR, L. B. (1990), *Understanding Significance Testing*, Sage Publications, Newbury Park, CA.

MOHRI, M. (1996), On some applications of finite-state automata theory to natural language processing, *Journal of Natural Language Engineering*, 2 : 61–80.

MOED, H. F. (2000), Bibliometric indicators reflect publication and management strategies, *Scientometrics*, 47 : 323–346.

MOED, H. F., VAN RAAN, A. F. J. (1988), Indicators of research performance: Applications in university research policy. In: A. F. J. VAN RAAN (Ed.), *Handbook of Quantitative Studies of Science and Technology*. Elsevier Science Publishers, Amsterdam, pp. 177–192.

MOED, H. F., VIRIENS, M. (1989), Possible inaccuracies occurring in citation analysis, *Journal of Information Science*, 15 : 95–117.

MOYA-ANEGÓN, F., VARGAS-QUESADA, B., HERRERO-SOLANA, V., CHINCHILLA-RODRÍGUEZ, Z., CORERA-ÁLVAREZ, E., MUNOZ-FERNANDEZ, F. J. (2004), A new technique for building maps of large scientific domains based on the cocitation of classes and categories, *Scientometrics*, 61 : 129–145.

MOYA-ANEGÓN, F., VARGAS-QUESADA, B., CHINCHILLA-RODRÍGUEZ, Z., CORERA-ÁLVAREZ, E., HERRERO-SOLANA, V., HERRERO-BOTE, V. (2003), SCImago: A proposal of integrated visual scientific information systems. In: *Proceedings of the 9th International Conference on Scientometrics & Informetrics (ISSI-2003)*.

NOYONS, E. C. M., MOED, H. F., LUWEL, M. (1999), Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study, *Journal of the American Society for Information Science*, 50 : 115–131.

PAO, M. L. (1989), Importance of quality data for bibliometric research. In: C. NIXON, L. PADGETT (Eds), *National Online Meeting. Proceedings*. Learned Information, Medford, NJ, pp. 321–327.

PAUMIER, S. (2003), *De la reconnaissance de formes linguistiques a l'analyse syntaxique*, Ph.D., Université de Marne-la-Vallée.

PITERNICK, A. B. (1982), Standardization of journal titles in databases (letter to the editor), *Journal of the American Society for Information Science*, 33 : 105.

RICE, R. E., BORGMAN, C. L., BEDNARSKI, D., HART, P. J. (1989), Journal-to-journal citation data: Issues of validity and reliability, *Scientometrics*, 15 : 257–282.

RINIA, E. J., DE LANGE, C., MOED, H. F. (1993), Measuring national output in physics: Delimitation problems, *Scientometrics*, 28 : 89–110.

ROCHE, E. (1993), *Analyse Syntaxique Transformationelle du Français par Transducteurs et Lexique-Grammaire*, PhD thesis, Université Paris, Paris.

ROCHE, E. (1996), Finite-state transducers: Parsing free and frozen sentences. In: A. Kornai (Ed.), *Proceedings of the ECAI 96 Workshop Extended Finite State Models of Language*. ECAI, pp. 52–57.

ROCHE, E., SCHABES, Y. (1995), Deterministic part-of-speech tagging with finite state transducers, *Computational Linguistics*, 21 : 227–253.

SHER, I. H., GARFIELD, E., ELIAS, A. W. (1966), Control and elimination of errors in ISI services, *Journal of Chemical Documentation*, 6 : 132–135.

SHRUM, W., MULLINS, N. (1988), Network analysis in the study of science and technology. In: A. F. J. VAN RAAN (Ed.), *Handbook of Quantitative Studies of Science and Technology.* Elsevier Science Publishers, Amsterdam, pp. 107–133.

SILBERZTEIN, M. (1993), *Dictionnaires Électroniques et Analyse Automatique de Textes: Le Systčme INTEX*, Masson, Paris.

SILBERZTEIN, M. (2000), INTEX: An FST toolbox, *Theoretical Computer Science*, 231 : 33–46.

STEFANIAK, B. (1987), Use of bibliographic data bases for scientometric studies, *Scientometrics*, 12 : 149–161.

THE THOMSON CORPORATION (2005), ISI Web of Science. Available from: http://isiknowledge.com (visited: 11/07/2005)

VAN DEN BERGHE, H., DE BRUIN, R. E., HOUBEN, J. A., KINT, A., LUWEL, M., SPRUYT, E., MOED, H. F. (1998), Bibliometric indicators of university research performance in Flanders, *Journal of the American Society for Information Science*, 49 : 59–67.

VAN RAAN, A. F. J. (1993), Advanced bibliometric methods to assess research performance and scientific development: Basis principles and recent practical applications, *Research Evaluation*, 3 : 151–166.

VAN RAAN, A. F. J. (1999), Avanced bibliometric methods for the evaluation of universities, *Scientometrics*, 45 : 417–423.

VAN RAAN, A. F. J. (2003), The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments, *Technikfolgenabschätzung -Theorie Und Praxis*, 12 : 20–29.

VAN RAAN, A. F. J. (2005), Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics*, 62 : 133–143.

VAN RIJSBERGEN, C. J. (1979), *Information Retrieval*, Butterworths, London.

WILLIAMS, M. E., LANNOM, L. (1981), Lack of standardization of the journal title data element in databases, *Journal of the American Society for Information Science*, 32 : 229–233.

Appendix

Table 4. Extract of the binary matrix for the uniform representation of university department data