Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 72, No. 2 (2007) 167–183
DOI: 10.1007/s11192-007-1589-z

Variations in content and format of ISI databases in their different versions: The case of the Science Citation Index in CD-ROM and the Web of Science

RODRIGO COSTAS,[a] ISABEL IRIBARREN-MAESTRO[b]

[a] *Centro de Información y Documentación Científica, CINDOC-CSIC, Madrid (Spain)*
[b] *Laboratorio de Estudios Métricos de Información (LEMI), Departamento de Biblioteconomía y Documentación, Universidad Carlos III de Madrid, Getafe, Madrid (Spain)*

The CD-ROM and web versions of the Science Citation Index databases are compared as to their content and format features. Several differences have been detected such as the use of different punctuation marks in both versions and a different organisation of author's affiliation data. These differences make automatic comparisons of ISI products difficult and they should be considered when matching both databases. Some recommendations to ensure more normalisation and reliability of data are pointed out.

Introduction

Since its creation in 1958, the Institute for Scientific Information (ISI) has developed different products, such as the multidisciplinary databases Science Citation Index (SCI), Social Science Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI). Nowadays, these databases include 8700 journals in 35 different languages (THOMSON, 2005a), and cover all the scientific disciplines. The main advantage of these databases, with respect to other bibliographic databases, is the data processing of bibliographic references of original articles. This fact allows us to establish links between different articles, thus fulfilling the initial idea of its creator,

Received December 19, 2005

Address for correspondence:
RODRIGO COSTAS
Centro de Información y Documentación Científica. CINDOC-CSIC
C/Joaquín Costa, 22, 28002, Madrid, Spain
E-mail: rodrigo.costas@cindoc.csic.es

0138–9130/US $ 20.00

Eugene Garfield, of creating citation networks in the different scientific fields through citation indexes (THRACKRAY & BROCK, 2000).

Besides the processing of bibliographical references, ISI adds to its databases statistical data from its covered journals (*Journal Citation Reports*, JCR), with bibliometric indicators such as the Impact Factor of every journal for every year.

Due to the amount of information covered by ISI databases, they are frequently used in bibliometric studies. However, they have been very criticized in different aspects, such as the questionable value of its citations, the usefulness of Impact Factor, the subject classification (SANZ et al., 2002; VOUTIER, 2004), and the database coverage, which show biases toward the Anglo-Saxon scope, basic sciences and articles (in detriment of other types of document) (MOED & VAN LEEUWEN, 1995; GARFIELD, 1996; GÓMEZ & BORDONS, 1996; ADAM, 2002).

Furthermore, there are more technical limitations, such as the lack of normalization of some fields (GOMEZ & GALBAN, 1986; FERNÁNDEZ et al., 1993; RUIZ PÉREZ et al., 2002), the lack of periodical updating of some versions such as CD-ROM, or the limitations of the newest version (ISI Web of Science) to download items. All these restrictions can affect the results of the bibliometric studies. Nevertheless, it is important to be aware that in spite of the scientometric foresight made by Garfield, these databases were developed from a bibliographic point of view, and therefore their information is more focused on this mission (MOED, 2002) than on being the object of bibliometric or scientometric studies.

Throughout ISI history, the format of its products has changed according to scientific and technological advances (paper, magnetic tape, floppy disk, CD-ROM, DVD and web format) (THOMSON, 2005b). Although these products can coexist in more than one medium, they sometimes give rise to confusion due to the differences observed between different editions. These differences must be taken into consideration when a bibliometric study is done, since they can determine the methodology that has to be used for the data treatment and for the interpretation of the results. Currently, two of the most important formats for obtaining data from ISI databases are CD-ROM (since 2004 SCI database also exists in DVD) and the platform Web of Science (WOS) (http://go5.isiknowledge.com). The coexistence of these different formats prompts us to talk about two different versions of the same database. The drawbacks and the benefits of each format are stated below, consider each before carrying out a bibliometric study.

CD-ROM

ISI publishes three CD-ROMs per year, one for each of its databases (SCI, SSCI and A&HCI), allowing researchers to know about the latest scientific publications in each area. Moreover, these databases are useful for finding articles mentioned in previous studies and doing interdisciplinary research (THOMSON, 2005a). Each CD-ROM includes the scientific papers published the previous year in the journals analysed by the ISI. This version has a number of advantages and limitations:

Advantages:

- There is no limit on the number of documents to be downloaded.
- Downloads can be executed in different formats, making it easier to export them to other computer programs.
- Data regarding the author's institutional affiliation are clearly identified in one field: "Address".

Limitations:

- For searches covering more than one year it is necessary to execute them in several CD-ROM, according to the period of time contained in each CD-ROM.
- Citations received by each document are not included.
- It covers 2000 journals less than the Web of Science.
- For searches by cited references, it is only possible to look for the first author, and it can only be done once per year.

Web of Science (WOS)

This version, developed by ISI in 2001, which has been considered the only method of cited reference searching (SEVINC, 2004), also has a number of advantages and limitations.

Advantages:

- Data is updated weekly, because of the electronic format.
- It is more exhaustive than the CD-ROM because it includes 2000 more journals.
- Searches covering different years can be done simultaneously.
- The interface is more user-friendly than that of the CD-ROM and allows us to create search strategies to find relevant literature through the references.
- It allows us to download data in different formats.
- For searches by cited references, it is possible to look for any author regardless of his/her place in the document (at least for "ISI documents")

Limitations:

- It is only possible to download 500 documents each time, and this fact reduces the manoeuvrability of the database.
- Authors' address data is split up in two different fields: "Reprint Address" (RP) and "Author Address" (C1).

Differences among ISI products have been analysed by WHITLEY (2002), while JAIN (2005) has commented that any analysis based on SCI should always mention the version used. SATYANARAYANA & JAIN (2004) pointed out that data obtained from the

CD-ROM version and from WOS can show different results when studying the same object; they recommend the use of the WOS version instead of the CD-ROM. However, many times it is necessary to update studies from the CD-ROM version with WOS data, for example, including the number of citations received by documents. The change of version or the need to update data from different versions of the same database means having to deal with different problems.

Similar downloads from different SCI versions show that neither the file of data nor the visualization mode coincide. This entails the need to make a detailed analysis in order to establish the formal differences between the two formats and to suggest solutions and strategies to surpass these problems.

Of special importance are the differences between WOS and CD-ROM in the author's affiliation data, since they are essential to bibliometric studies allowing us to carry out different kinds of studies (macro, meso and micro) about countries, regions, institutions, etc.

Objectives

In this work, three main objectives have been identified:

- To analyse the main differences between the two more common versions for accessing the ISI databases: CD-ROM and WOS.
- To analyse the characteristics of the two fields related to author's author data in the WOS: field RP and C1.
- To establish recommendations that allow the users of these databases to use their data in order to collate both databases correctly.

Methodology

In order to carry out this study, scientific publications of researchers from Natural Resources Area of the Spanish CSIC during the years 1994–2001 were collected. Data were downloaded simultaneously from CD-ROM and WOS of the SCI database [in the case of CD-ROM, the search was done in "address field"; in the case of WOS, the search was carried out in "Address Field", because the database automatically looks in fields "address" and "RP"]. Two sets of 3301 matching documents from the two versions (CD-ROM and WOS) were obtained; CD-ROM was the limit because, as has previously been mentioned, it presents more restrictions than WOS.

Both sets of documents were treated in a relational database following the methodology proposed by FERNÁNDEZ et al. (1993) making data management and analysis easier.

Moreover, both data sets (CD-ROM and WOS) were linked document by document through paper title matching, journal title matching, author name matching, and so forth; documents not linked automatically were linked by hand.

Finally, both sets were compared and analysed, detecting every difference among fields with the same type of information (Author field, Journal Title field, Author address fields, etc.)

Results

The results regarding the differences between both data sets in main data fields are presented below. The fields that have been analysed are Paper title, Journal title, Publication data (year of publication, pages, number of volume and number of issue), Author field and Author's affiliation fields.

Title field

The differences in Title field were analysed. This is a main field from a bibliographic point of view, vital to identify a document.

A total of 37.86% of documents presented exact textual coincidence of Titles between both data sets. By deleting ("-") hyphen characters and preceding and subsequent blank spaces (" "), 969 records were also coincident, representing 29.35% of documents.

In the remaining documents, several punctuation marks such as colon (":"), semicolon (";"), dot (".") and comma (","), were deleted detecting 518 new coincident documents (15.69%).

Another particular difference is that in WOS, numbers are written in letters while in CD-ROM they are written with numbers. By replacing in WOS written values from one to ten with numbers, 96 new coincident documents were detected, representing 2.91% of documents.

Finally, problems with blank spaces were resolved by eliminating all blank spaces in title field in both data sets. With this correction, 180 coincident documents (5.45%) were detected.

After format and punctuation marks depuration, 91.26% of documents matched in Title data. Table 1 shows the different circumstances identified.

As can be observed in Table 1, 8.72% of documents were not matched between both data sets due to different problems such us using abbreviations in one data set and full names in the other, typographic mistakes in either of the two data sets, etc.

Table 1. Matching title fields

Type of difference	No. Documents	%
Total matching	1250	37.87
Differences by hyphen "-"	969	29.35
Differences by punctuation marks ";,:."	518	15.69
Differences by numbers	96	2.91
Differences by blank spaces	180	5.45
No matching titles	288	8.72
Total number of documents	3301	100

Journal Title field

After matching the full title of journals between both data sets, 3257 documents (98.67%) present an exact matching while 23 of remaining documents (0.70%) present differences due to the presence-absence of hyphens ("-") in journal title. The last 21 remaining documents (0.64%) present differences related with Journals subtitles (both data sets include sometimes subtitles) and typographic mistakes (Table 2).

Table 2. Matching journal data

Type of difference	No. Documents	%
Total Matching	3257	98.67
Differences by hyphen "-"	23	0.70
No matching titles	21	0.64
Total No. of documents	3301	100

Publication data fields (Year, Volume, Issue and Pages)

A total matching was found in 100% of documents between both data sets.

Author Field

Number of authors. The number of signing authors of every data set were analyzed, and it has been observed that 100% of documents have the same number of authors in both datasets.

Author matching. Comparing author fields of both data sets, differences were detected due to punctuation marks: commas (","), hyphens ("-") and blank spaces among surnames. It must be taken into account that sometimes WOS includes hyphens and blank spaces among surnames and always includes commas before first name initials. On the contrary, CD-ROM presents only one single string of characters for surnames and separates surnames from first name initials by a hyphen (i.e. "Garcia Perez, P" or "Garcia-Perez, P" in WOS, "Garciaperez-P" in CD-ROM). These detected differences are shown in Table 3.

The total number of compared authors was 12973.

Table 3. Differences in Author field data

Type of difference	No. string of Authors	%
Differences by comma ","	11890	91.65
Differences by hyphen "-"	873	6.73
Differences by blank spaces ""	210	1.62
Total No. strings of authors	12973	100

Author's institutional affiliation fields

In this study two different analyses are developed. On the one hand, the particularities of affiliation data coming from WOS are analysed: WOS presents affiliation data in two different fields C1 and RP. On the other hand, affiliation fields in CD-ROM and WOS are compared.

Web of Science analysis. Differences between C1 and RP fields. As was said before, the Web of Science databases present author's affiliation data in two different fields: a) RP (Reprint Address): this field contains data from the reprint author and his/her address; the aim of this field is to provide a person and an address to whom comments or communications about the paper can be sent. b) C1 (Author address), in which all institutional addresses of all researchers are included. According to ISI communications (personal communication from ISI) data gathered in these fields is extracted directly from the journals, and C1 is supposed to be the field where all institutional addresses of all researchers are included (even RP address), whereas RP field only includes just one author with his/her postal address. However, it has been observed that this situation does not always happen and sometimes the address included in RP is not included in C1 and in these cases the two fields are complementary.

In Table 4 all possible data combinations between fields RP and C1 are shown in data extracted from WOS. Seven different situations of combination have been detected.

Table 4. Possible situations concerning C1 and RP data from Web of Science

Situation	RP	C1	No. Documents	%
1	0 address	1 address	94	2.85
2	0 address	Several addresses	195	5.91
3	1 address	0 address	395	11.97
4	1 address	1 address (equal to RP)	552	16.72
5	1 address	1 address (different to RP)	458	13.87
6	1 address	Several addresses (one of them equal to RP)	1249	37.84
7	1 address	Several addresses (different to RP)	358	10.85
Total number of docs.			3301	

In Situation 1 (S1), documents with no data in RP and just one single address in C1 are included (as can be seen in Example 1). This situation was observed in 2.85% of documents.

Example 1. Example of one document with RP empty and just one affiliation address in C1.

```
AU Bastir, M
   Rosas, A
TI Thin plate splines analysis of human craniofacial sexual dimorphism.
SO AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY
LA English
DT Meeting Abstract
C1 Museo Nacl Ciencias Nat, Dept Paleobiol, Madrid, Spain.
NR 0
```

Situation 2 (S2) comprises of documents with no RP and several addresses in C1, this has been observed in 5.91% of the documents.

Situation 3 (S3) refers to documents with no data in C1 and one affiliation address in RP, as can be seen in Example 2. This situation exists in 11.97% of documents, in which the RP field add information not presented in C1.

Example 2. Example of one document with C1 empty and one affiliation address in RP.

```
AU Prego, R
TI Nitrogen interchanges generated by biogeochemical
  Processes in a Galician ria
SO MARINE CHEMISTRY
LA English
DT Article
ID SEA-WATER; CARBON; FLOWS; VIGO
RP Prego, R, CSIC,Inst Invest Marinas, Eduardo Cabello 6,E-36208
   Vigo, Spain.
CR ALVAREZ G, 1980, THESIS U SANTIAGO CO
```

Situation 4 (S4) refers to documents with one address in RP and just one address in C1, the address being the same in both cases (see Example 3).

Example 3. Example of one document with one address in RP and one address in C1, the address being the same.

```
AU Ferrer, M
   Harte, M
TI Habitat selection by immature Spanish imperial eagles during the
  dispersal period
SO JOURNAL OF APPLIED ECOLOGY
LA English
DT Article
DE Aquila adalberti; Donana National Park; endangered species; management;
   recovery plans
ID NEST-SITE SELECTION; AQUILA-ADALBERTI; JUVENILE DISPERSAL;
  MANAGEMENT; POPULATION; DENSITY; SUCCESS; SPAIN
C1 CSIC, Estac Biol Donana, Seville 41013, Spain.
RP Ferrer, M, CSIC, Estac Biol Donana, Avd Maria Luisa,Pabellon Peru,
   Seville 41013, Spain.
TC 9
```

Example 3 shows a document which presents “the same” address in RP and in C1. However, the addresses are not textually equal although they refer to the same institution (“Estac Biol Donana”) because RP address includes postal data about the avenue where the centre is located (“Avd Maria Luisa,Pabellon Peru”), which is not included in C1.

Situation 5 (S5) appears in those documents with just one address in C1 and one address in RP, being different addresses (see Example 4). This situation occurs in 13.87% of the documents. In this situation it is clear that C1 and RP provide complementary information.

Example 4. Example of one document with one address in RP and one address in C1, but they are different.

AU Salto, R
Delgado, A
Michan, C
Marques, S
Ramos, JL
TI Modulation of the function of the signal receptor domain of XylR, a member of a family of prokaryotic enhancer-like positive regulators
SO JOURNAL OF BACTERIOLOGY
LA English
DT Article
C1 CSIC, Estac Expt del Zaidin, Dept Biochem Mol & Cellular Biol Plants, E-18008 Granada, Spain.
RP Salto, R, Univ Granada, Sch Pharm, Dept Biochem & Mol Biol, E-18071 Granada, Spain.
EM rsalto@goliat.ugr.es

Situation 6 (S6) is the most common of all situations, being present in 37.84% of the documents. It refers to documents with one address in RP and several addresses in C1. Addresses in C1 include the address from RP (see Example 5).

Example 5. Example of one document with one address in RP included among the several addresses from C1.

AU Avila, A
Alarcon, M
Queralt, I
TI The chemical composition of dust transported in red rains - Its contribution to the biogeochemical cycle of a Holm oak forest in Catalonia (Spain)
SO ATMOSPHERIC ENVIRONMENT
LA English
DT Article
C1 Univ Autonoma Barcelona, Ctr Recerca Ecol & Aplicac Forestals, Bellaterra 08193, Barcelona, Spain.
Univ Politecn Catalunya, Dept Fis & Engn Nucl, Barcelona, Spain.
CSIC, Inst Cienciea Terra Jaume Almera, E-08028 Barcelona, Spain.
RP Avila, A, Univ Autonoma Barcelona, Ctr Recerca Ecol & Aplicac Forestals, Bellaterra 08193, Barcelona, Spain.
TC 21

When S6 occurs, it is very common that the address from RP which appears in C1 is the first one from C1, however, several cases were detected in which the address from RP appears in second place in C1.

Finally, situation 7 (S7) involves documents with one address in RP and several addresses in C1, but the address from RP is not included in C1, this situation is present in 10.85% of documents.

- Variations in RP data.

In those cases in which RP and C1 have an "equal" address (for example S4 and S6) two situations of equalities are found: on the one hand an exact textual matching between the coincident address in RP and C1 can be found; while on the other hand, this matching can only be approximate in other documents, the coincident address referring to the same location but RP address including postal data such as streets, avenues, PO Box, etc. (see the previous Example 3). This latter situation impedes an automatic matching of addresses.

However, it has been observed that addresses from RP present postal data in the third position before the end of the field very frequently and, less frequently, in the fourth position before the end of the field. "Position" refers to every part of the field delimited by comma-blank space (", "). We have denominated "Postal Data Type 1" to those addresses whose postal data appears in the third from last position before the end of the field while "Postal Data Type 2" refers to addresses with postal data in the fourth from last position.

As can be seen in Table 5, documents from Situations 4 and 6 (S4 and S6) have been analysed because they are the only situations in which RP is included in C1 (it occurs in 54.56% of documents). When this happens, the address from RP exactly matches with C1 in 363 documents (11%) (see the previous Example 5).

Table 5. Differences between RP and C1 when RP is included in C1 (No. of documents)

Type of differences	No. of Documents S4	No. of Documents S6	TOTAL	% total Documents (*)
Exact matching	70	293	363	11.00
Postal Data type 1	431	843	1274	38.59
Postal Data type 2	50	100	150	4.54
Other differences	1	13	14	0.42
TOTAL docs.	552	1249	1801	54.56

(*) % in relation to the total number of documents (3301).

In 1274 documents (38.59%) RP address has "Postal Data Type 1" and doesn't match exactly with its coincident address in C1 (see Example 6).

Example 6. Example of one document with RP and C1 coincidence, but with Postal Data Type 1.

C1 **CSIC, Inst Ciencia Marinas de Andalucia, Cadiz 11510, Spain.**
Univ Cadiz, Fac Ciencias Mar, Dept Biol Anim Vegetal & Ecol, Cadiz, Spain.
CSIC, Ctr Invest Biol, Madrid, Spain.
Univ Malaga, Fac Ciencias, Dept Microbiol, E-29071 Malaga, Spain.
RP Sarasquete, C, CSIC, Inst Ciencia Marinas de Andalucia, **Poligono Rio San Pedro,Apdo Oficial**, Cadiz 11510, Spain.

As can be seen in Example 6, an address from RP and C1 refers to the same affiliation but in the address from RP there are postal data in the third from last position which makes it difficult to match automatically the address.

Concerning "Postal Data Type 2" they are present in 4.54% of documents, see an example in Example 7.

Example 7. Example of one document with RP and C1 coincidence, but with Postal Data Type 2.

C1 **CSIC, Ctr Estudios Avanzados Blanes, Blanes 1101, Girona, Spain.**
Int Inst Infrastruct Hydraul & Environm Engn, NL-2601 DA Delft, Netherlands.
NIOO, Ctr Estuarine & Coastal Engn, NL-4401 NT Yerseke, Netherlands.
Univ Philippines, Inst Marine Sci, Diliman 1101, Quezon City, Philippines.
RP Duarte, CM, CSIC, Ctr Estudios Avanzados Blanes, **Camiino Santa Barbara S-N**, Blanes 1101, Girona, Spain

Finally, other differences between RP and C1 have been found, for example the inclusion of the names of academic departments instead of postal data, they are only present in 0.42% of documents.

- Analysis of the signing place of RP author in the document

The places of signing of authors present in RP have been analysed, with the aim of determining whether the Reprint Author corresponds with the first place signing author of the document. As can be seen in Table 6, in this case, 2841 documents were analysed because 289 documents with the RP field empty and 171 documents with only one author were discarded.

Table 6. Signing place of RP authors in documents

Signing place	No. Documents	%
First	2492	87.72
Last	201	7.07
Middle	148	5.21
	2841	

Table 6 shows that 87.72% of documents have as first author the one who is in the RP field. However, in 7.07% of documents the author from RP is signing in the last place and in 5.21% of documents he/she is in an intermediate place.

Differences between CD-ROM and WOS in the author's affiliation field. After detecting differences between fields RP and C1 from WOS, differences in affiliation fields between WOS and CD-ROM have also been analysed.

A very common difference that has been detected consists of some addresses from WOS not having any blank spaces after a comma, while in CD-ROM address have blank spaces after commas. Another very frequent difference is that in WOS a hyphen "-" appears inside the postal code (see previous Examples 2, 4 or 5) while in CD-ROM this hyphen doesn't appear. The problem is that originally data from CD-ROM contains hyphens among words, and in a previous cleansing process all hyphens are deleted, that is the reason why postal codes do not have hyphens.

Table 7 summarizes the main characteristics in the comparison of address between WOS and CD-ROM.

Table 7. Main differences between CD-ROM address and WOS address

Difference WOS-CD-ROM	No. Addresses	% of Addresses	No. Documents	% Documents
Exact matching	1494	20.64	978	29.63
Differences in commas, blank spaces and/or hyphens	5334	73.68	2907	88.06
Transcription mistakes	411	5.68	334	10.12

Note: Since a document may have more than one address, the sum of % of documents is higher than 100. The % of addresses is calculated based on the 7239 addresses of the CD-ROM.

As can be observed in Table 7, addresses are exactly equal in WOS and CD-ROM in 1494 cases which comprises of 978 documents (29.63%). 5334 addresses present differences due to commas, blank spaces and/or hyphens, included in 2907 documents (88.06% of documents). There are also differences that can be considered as transcription mistakes, for example "CASTELLO DE PLANA" in WOS and "CASTELLO DE LA PLANA" in CD-ROM, "SEVILLE" in WOS and "SEVILLA" in CD-ROM, etc.

C1 and RP fields from WOS have been compared with the Address field from CD-ROM to find out which addresses from WOS are presented in CD-ROM. There are four possibilities:

1. Address in CD-ROM is the same as RP field (Only RP).
2. Addresses in CD-ROM are the same as C1 field (Only C1).
3. Addresses in CD-ROM are the same as both fields (C1 and RP), including the address from RP and the rest from C1.
4. Addresses in CD-ROM are the same as fields RP and C1 simultaneously. This happens when the address from RP is textually identical to the same in C1.

Only documents were taken into account where the address matched exactly or before a comma and/or a hyphen cleansing in order to do this analysis. The transcription mistakes were not analysed because it is difficult to assign the address to RP or to C1. A total of 2849 documents were studied.

The main results are summarized in Table 8 itemised by type of situation existing between RP and C1 analysed previously to explain every case better.

Table 8. Origin of CD-ROM's affiliation data in relation with WOS's fields (RP and C1)

Presence/Absence of RP/C1 data in CD-ROM	Total	%	S1	S2	S3	S4 RP=C1	S5 RP≠C1	S6 RP=C1	S7 RP≠C1
Only RP	776	27.24	0	0	389	385	2	0	0
Only C1	586	20.57	91	160	0	82	37	154	62
C1 and RP	1208	42.40	0	0	0	1	368	609	230
C1=RP	279	9.79	0	0	0	60	0	219	0
Total	2849		91	160	389	528	407	982	292

Looking at Table 8 it can be observed that 776 documents in CD-ROM present the address from RP field in WOS. This happens mainly in S3 and S4, in S3 C1 it is empty and in S4 C1 and RP are equal, but in this latter case CD-ROM prefers data from RP (see Example 8).

Example 8. Example of one WOS address with only RP vs. the same address in the same document in CD-ROM.

WOS:

RP Sitjabobadilla, A, **CSIC, Inst Acuicultura Torre Sal,E-12595 Ribera de Cabanes, Spain**.

CD-ROM:

Addresses:
CSIC, INST ACUICULTURA TORRE SAL, E-12595 RIBERA-DE-CABANES, SPAIN

It is observed that 586 CD-ROM documents show the C1 address, which happens mainly in S2 because the RP is empty, and in the S6 because RP and C1 are coincident (see Example 9).

Example 9. Example of one document with C1 address in CD-ROM.

WOS:

C1 **CSIC, Inst Jaume Almera, Barcelona 08028, Spain.**
Univ Complutense, Fac Fis, Dept Fis Aplicada 3, Madrid 28040, Spain.
RP Artus, L, **CSIC, Inst Jaume Almera, Lluis Sole & Sabaris SN, Barcelona 08028, Spain.**

CD-ROM:

Addresses:
CSIC, INST JAUME ALMERA, BARCELONA 08028, SPAIN
UNIV-COMPLUTENSE, FAC FIS, DEPT FIS APLICADA 3, MADRID 28040, SPAIN

A total number of 1208 documents in the CD-ROM have a combination of data from C1 and RP; it means that data in CD-ROM come from both C1 and RP, and when the address in C1 and RP is the same (but not textually the same), RP address is preferable to C1 to be included in CD-ROM. See Example 10.

Example 10. Example of a document with the same address in RP and C1, the RP address being in CD-ROM.

WOS:

C1 CSIC, **Ctr Ciencias Medioambientales, E-28006 Madrid, Spain.**
Univ La Laguna, Fac Biol, UDI Fitopatol, Tenerife, Spain.
CSIC, Inst Prod Nat & Agrobiol, Tenerife 38206, Spain.
RP Gonzalez-Coloma, A, CSIC, **Ctr Ciencias Medioambientales, Serrano 115-Dpdo, E-28006 Madrid, Spain.**

CD-ROM:

Addresses:

CSIC, CTR CIENCIAS MEDIOAMBIENTALES, SERRANO 115-DPDO, E-28006 MADRID, SPAIN
UNIV-LA-LAGUNA, FAC BIOL, UDI FITOPATOL, TENERIFE, SPAIN
CSIC, INST PROD NAT & AGROBIOL, TENERIFE 38206, SPAIN

Finally, 279 documents show exact matching of addresses in both C1 and RP. This is the reason why it is impossible to know whether the address in CD-ROM comes from C1 or RP. See Example 11.

Example 11. Example of an address with textual matching between RP and C1 in WOS and in CD-ROM.

WOS:

C1 **Univ Complutense, Fac Ciencias Biol, Dept Biol Anim Fisiol 2, E-28040 Madrid, Spain**.
CSIC, Museo Ciencias Nat, E-28006 Madrid, Spain.
RP Puerta, M, **Univ Complutense, Fac Ciencias Biol, Dept Biol Anim Fisiol 2, E-28040 Madrid, Spain.**

CD-ROM:

Addresses:

UNIV-COMPLUTENSE, FAC CIENCIAS BIOL, DEPT BIOL ANIM FISIOL 2, E-28040 MADRID, SPAIN
CSIC, MUSEO CIENCIAS NAT, E-28006 MADRID, SPAIN

Conclusions

About bibliographic data

With regard to the basic bibliographic fields analysed in this study, a high coincidence between CD-ROM and WOS is found: by 100% of publication data, by 98% of journal titles, by 92% of number of authors per document (after removing commas, because if commas were taken into account, the percentage would be 0). But this coincidence decreases to 30% in the field related to title of articles.

Two types of differences were identified: firstly, there are some differences that can be named as “minor”, because its correction can be easily systematized. Some of them are punctuation marks (commas, hyphens, etc.) or blank spaces (in the case of title of document, this difference is shown in about 50% of documents, and in the case of journal title it is in 0.70% of documents). These minor differences make the automatic identification of records difficult, but they can be easily surpassed if we know about it in advance. Secondly, there are some more complex differences, which should be reviewed carefully such as random typographic differences. About 11% of the documents show this difference in the field of document title and the 0.64% in the journal title field.

About author's affiliation data

The Web of Science analysis: differences between C1 and RP fields. Most of the documents (63.31%) do not include any new information in the RP field. If this information is included, it is the same as the information covered by the C1 field in the Web of Science. However, by 36.69% of documents, the information of RP is additional to the C1 field, which makes us aware of the importance of taking into account both fields in the development of a bibliometric study so that the information regarding author's affiliation data is not lost.

The results about the similarities among textual structures also show the possibility of standardising automatically the address information from both C1 and RP (taking into account the mentioned information about postal code 1 and 2).

Up to 87% of items in which there are more than one author and the RP field contains information, the author from RP field is the same as the first author of the document.

What is not known is the reason why author's affiliation data show the seven situations described in this study related to C1 and RP fields.

Differences between CD-ROM and WOS in author's affiliation data. 30% of the documents show addresses absolutely coincident in both databases. 88% of the documents have addresses which are coincident after having an automatic treatment of commas and hyphens. Only 14% of the documents show addresses which should be analysed in detail.

Comparing the address fields from the WOS (C1 and RP) with the address field from the CD-ROM, it is possible to verify that the CD-ROM has preference to RP information.

Recommendations

Generally speaking, it is not possible to talk about a complete textual coincidence, this coincidence being very small in important fields such as document title (30%), author names (0% owing to the use of commas in the WOS) or author affiliation data (30%, on account of the separation of this information in two fields: C1 and RP). Because of the differences between WOS and CD-ROM many times users had to check the data by hand.

An algorithm with several steps about data management is outlined in order to automatically make the data normalisation from ISI CD-ROM and WOS versions.

1. Matching documents from the two data sets (WOS and CD-ROM).

1.1 Puntuaction marks (",", ";", ":" and "."), hyphens and blank spaces should be removed from Article Title, Journal Title and Author fields in both databases (original data from these fields should be saved in different fields in order to keep them for further analysis).

1.2 Matching for Title, Journal Title, Publication Year and Publication data Fields. A large number of documents should match in this step; the accuracy of this step is good enough to avoid more revisions.

1.3 Matching for Journal Title, First author, Publication Year and Publication data of remaining documents. In this step, no mistakes should be found; however, it would be advisable to briefly revise the matchings.

1.4 Matching for Publication Year, First author and Publication data of remaining documents. A careful revision of matchings should be done.

1.5 Remaining documents should be handled manually to check the mistakes.

2. Affiliation Data from WOS (C1 and RP cleansing).

2.1 In WOS RP field, author name should be noted separately, keeping only the affiliation address.

2.2 Remove puntuaction marks and blank spaces from both fields (RP and C1), except commas.

2.3 Combine both fields (RP and C1) and remove duplicates.

2.4 Separate the Postal Data Type 1 (data between third and second from last comma) in remaining RP addresses. Remove duplicates with C1 addresses.

2.5 Separate the Postal Data Type 2 (data between fourth and third from last comma) in remaining RP addresses. Remove duplicates with C1 addresses.

2.6 If after this process, matching with addresses from CD-ROM is required; punctuation marks should be deleted in both datasets and matched. Remaining mistakes should be checked by hand.

Despite this algorithm, it is important to bear in mind that it is very difficult to solve 100% of the normalisation problems; therefore, manual and further revisions should be done in order to avoid mistakes and to complete the normalisation (problems such as different transcriptions of toponyms and numbers should be carefully revised).

Finally, it would be strongly recommended that ISI normalise its data more conscientiously, making greater compatibility among their different formats possible.

*

This work has been done thanks to a CSIC I3P scholarship and a University Carlos III Foundation scholarship. We would like to thank to our supervisors Maria Bordons and Elías Sanz-Casado for their advice and comments throughout the development of this study, and we would like to thank Paula Alonso, Maria Varela and Angie Moore for their help in the translation of this paper.

References

ADAM, D. (2002), The counting house, *Nature*, 415 : 726–729.

FERNÁNDEZ, M. T., CABRERO, A., ZULUETA, M. A., GÓMEZ, I. (1993), Constructing a relational database for bibliometric analysis, *Research Evaluation,* 3 (1) : 55–62.

GARFIELD, E. (1996), How can impact factors be improved?, *British Medical Journal,* 313 : 411–413.

GÓMEZ, I., BORDONS, M. (1996), Limitaciones en el uso de los indicadores bibliométricos para la evaluación científica, *Política científica*, 46 : 21–27.

GÓMEZ, I., GALBÁN, C. (1986), Lack of standardisation in the corporate source field of different databases, *10th International Online Information Meeting. London, 2–4, Dec*, Learned Information, Oxford, pp. 335–352.

JAIN, N. C. (2005), Scopus™ has wider scope than Science Citation Index, *Current Science,* 88 (3) : 331.

MOED, H. F., VAN LEEUWEN, TH. N. (1995), Improving the accuracy of institute for scientific information's journal impact factors, *Journal of the American Society for Information Science*, 46 (6) : 461–467.

MOED, H. F. (2002), The impact-factors debate: the ISI's uses and limits, *Nature,* 415 : 731–732.

RUÍZ PÉREZ, R., DELGADO LÓPEZ-COZAR, E., JIMÉNEZ-CONTRERAS, E. (2002), Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies, *Journal of Medical Library Association*, 90 (4) : 411–430.

SANZ CASADO, E., MARTÍN MORENO, C., MAURA, M., RODRÍGUEZ, B., GARCÍA ZORITA, C., LASCURAIN SÁNCHEZ, M. L. (2002), Análisis de la interdisciplinariedad de los investigadores puertorriqueños en ciencias químicas, durante el período 1992-1999, *Revista Española De Documentación Científica*, 25 (4) : 421–432.

SATYANARAYANA, K., JAIN, N. C. (2004), Web of science: measuring and assessing science beyond SCI, *Current Science*, 86 (5) : 627–629.

SEVINC, A. (2004), Web of Science: a unique method of cited reference searching, *Journal of the National Medical Association*, 96 (7) : 980–983.

THACKRAY, A., BROCK, D. C. (2000). Eugene Garfield: History, Scientific Information, and Chemical Endeavor. In: B. CRONIN, H. BARSKY ATKINS (Eds), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield.* Information Today, Inc., New Jersey, pp. 11–25.

THOMSON (2005a), *Thomson Scientific: Products & Services* [Web Page], Available: http://scientific.thomson.com/products/

THOMSON (2005b), *ISI®Timeline* [Web Page], Available: http://www.isinet.com/aboutus/timeline/

VOUTIER, C. (2004), Web of Science: index not as useful as it appear, *Journal of the National Medical Association,* 96 (9) : 1240–1241.

WHITLEY, K. M. (2002), Analysis of SciFinder Scholar and Web of Science Citation Searches, *Journal of the American Society for Information Science and Technology*, 53 (14) : 1210–1215.