INTRODUCTION TO BIBLIOMETRIC DATA SOURCES

Wolfgang Glänzel

Centre for R&D Monitoring and Dept MSI, KU Leuven
MTAK, Dept. Science Policy & Scientometrics, Budapest, Hungary

(*In collaboration with Juan Gorraiz, University of Vienna, Library and Archive Services, for esss*)

ECOOM

Structure

Introduction

Data sources of bibliometric/scientometric research and technology are bibliographies and bibliographic databases. In principle, any appropriate and sufficiently large publication list could be used as data source.

However, for comparative studies standardised record formats using the same standard for both the units of analysis and the reference units are needed.

Introduction

Data sources of bibliometric/scientometric research and technology are bibliographies and bibliographic databases. In principle, any appropriate and sufficiently large publication list could be used as data source.

However, for comparative studies standardised record formats using the same standard for both the units of analysis and the reference units are needed.

Bibliometric analyses are therefore preferably based on large *specialised or multidisciplinary* bibliographic databases.

Data requirements for bibliometrics

Bibliographic databases are originally not designed for bibliometric use. Their original design was documentation and information, their most important feature was to support the retrieval of information.

Data requirements for bibliometrics

Bibliographic databases are originally not designed for bibliometric use. Their original design was documentation and information, their most important feature was to support the retrieval of information.

Bibliometrics has specific requirements that are not always met in an adequate manner by bibliographic databases.

Data requirements for bibliometrics

Bibliographic databases are originally not designed for bibliometric use. Their original design was documentation and information, their most important feature was to support the retrieval of information.

Bibliometrics has specific requirements that are not always met in an adequate manner by bibliographic databases.

Standard bibliometrics mainly uses databases that cover journal literature. At present non-serial proceedings literature mark the borderline of large-scale bibliometric use.

Indispensable information for bibliometrics

- unambiguous bibliographic description of indexed items (precondition)
- all-author recording
- all-address recording
- descriptors (keywords/subject headings)
- publication type information
- complete reference list (for citation analysis)
- document type information
- title and abstract (for text mining)
- acknowledgement (for funding information, sub-authorship)
- Open Access information

☞ Subject classification is often based on journal assignment

☞ Author identifiers (with reservations)

The "ISI Databases" – (historical view)

The databases of the *Institute for Scientific Information* (ISI - now part of Clarivate Analytics), above all, the Science Citation Index (SCI) and its extended version, the Science Citation Index Expanded (SCIE) have become a generally accepted basic source for bibliometric analyses.
It covers peer-reviewed scholarly journals in the life, natural and applied sciences.

The "ISI Databases" – (historical view)

The databases of the *Institute for Scientific Information* (ISI - now part of Clarivate Analytics), above all, the Science Citation Index (SCI) and its extended version, the Science Citation Index Expanded (SCIE) have become a generally accepted basic source for bibliometric analyses.
It covers peer-reviewed scholarly journals in the life, natural and applied sciences.

The SCI was actually the first large multidisciplinary bibliographic database that allowed advanced bibliometric studies. It was the first database the unique features of which met all basic requirements of bibliometric technology.

- Multidisciplinarity. All research fields in the life sciences, natural sciences, mathematics and engineering are represented.

The "ISI Databases" – (historical view)

- Selectiveness. Periodicals covered by SCI are chosen on the basis of quantitative criteria (impact measurements), and the selection is generally reinforced by expert opinion.
- Full coverage. All papers published in periodicals covered by the SCI are recorded.
- Completeness of addresses. The addresses of all authors are indicated, allowing analyses of scientific collaboration and the application of full publication counting schemes.
- Bibliographical references. Together with each document their references are processed. Redefining references as sources makes it possible to analyse citation patterns and to construct citation indicators.
- Availability. Data have been first provided in print but later on also in electronic form (first on magnetic tapes, later on other data storage devices and on-line).

The ISI SCI (historical view)

The SCI database is the oldest large multidisciplinary citation database. It was created in the early 1960s. It was first available in print. Electronic versions followed soon.
The main components of this edition were the following three indexes:

- Source Index
- Citation Index
- Permutation Index

The ISI SCI (historical view)

The SCI database is the oldest large multidisciplinary citation database. It was created in the early 1960s. It was first available in print. Electronic versions followed soon.
The main components of this edition were the following three indexes:

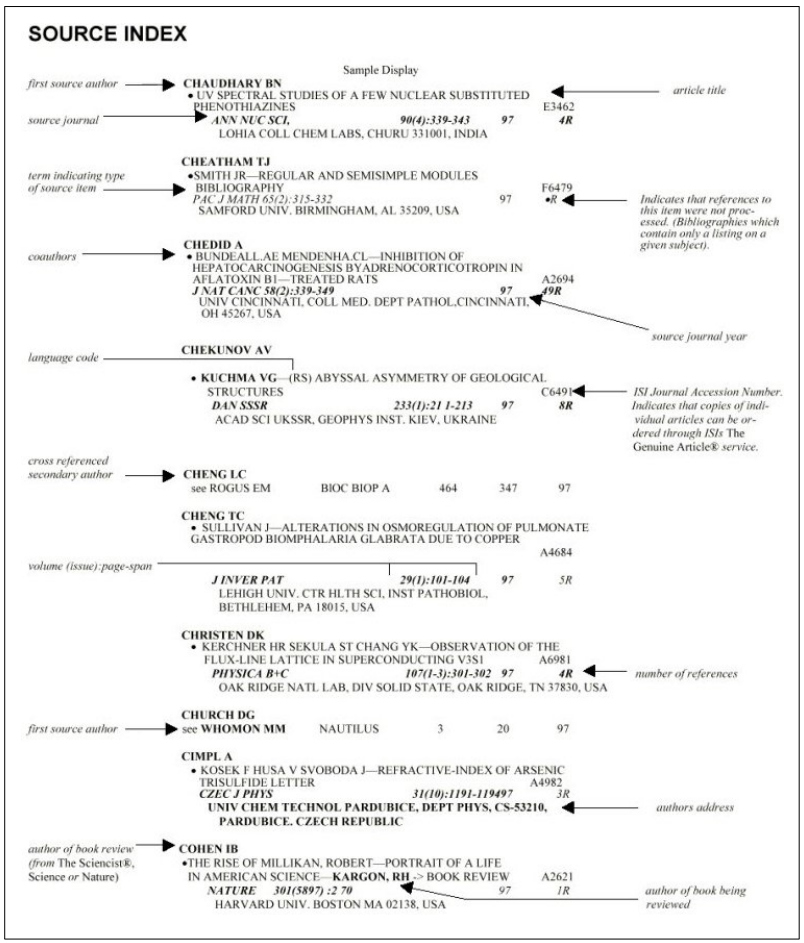
- Source Index
- Citation Index
- Permutation Index

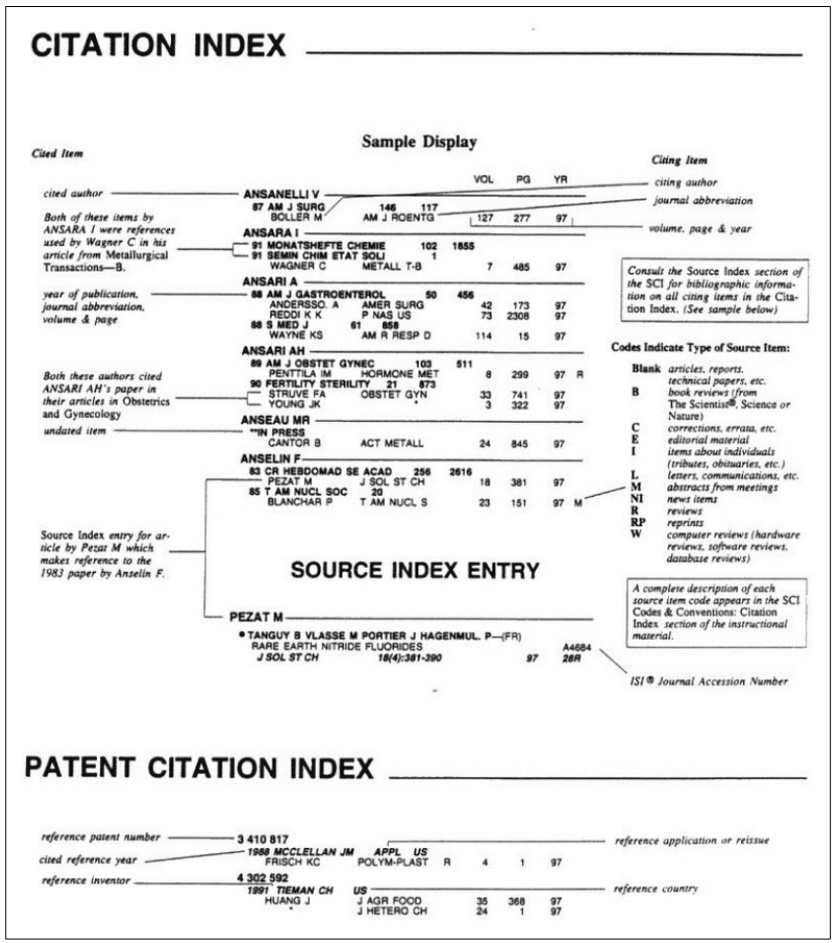
Each index covers the same material but indexes it differently. There is a large range of search options the SCI offered in different ways:

- by cited author and cited work or by cited patent (Citation Index)
- by source author (Source Index) or by source organization (Corporate Index, a section of the source index)
- by title words (Permuterm Subject Index)

The "ISI Databases" – (historical view)

The ISI SCI Print Edition





The Web of Science

At present, Clarivate Analytics *Web of Science Core Collection* (WoS) comprises the following products.

- Science Citation Index Expanded (SCIE)
- Social Sciences Citation Index (SSCI)
- Arts & Humanities Citation Index (A&HCI)
- Conference Proceedings Citation Index- Science (CPCI-S)
- Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH)
- Book Citation Index (BKCI)
- Emerging Sources Citation Index (ESCI)

- Journal Citation Reports (JCR) – now part of Clarivate Analytics InCites

Some limitations of Proceedings and Book citation indices from the bibliometric viewpoint:

- *Proceedings databases*
  - Sometimes missing or incomplete address information for co-authors.
  - Sometimes only first author information provided.
  - Not always unified conference information in conference series.
  - Sometimes missing proceedings volumes in conference series.
  - Citations less reliable than in journal databases.
- *Book Citation Index*
  - The absence of affiliation data in BKCI (Gorraiz et al., 2013)
  - The low share of BKCI indexed items with references data (Chi, 2014)
  - BKCI lacks a clear distinction of document types due to the different forms of book literature.
  - Citations to books and their individual chapters not systematically handled.

The Web of Science

The electronic versions provide many additional tools that go far beyond traditional information retrieval and are closely connected to bibliometrics:

- Related records,
- Analyse tools,
- Citation reports including elementary measures and indicators (mean citation rates, h-index, citation networks etc.)

The Web of Science

The electronic versions provide many additional tools that go far beyond traditional information retrieval and are closely connected to bibliometrics:

- Related records,
- Analyse tools,
- Citation reports including elementary measures and indicators (mean citation rates, h-index, citation networks etc.)

Clarivate Analytics *Web of Science* hosts along with the WoS other products such as the Chinese Science Citation Database, Current Contents, Derwent Innovations Index and a range of disciplinary databases such as MEDLINE, CAB Abstracts, Inspec and others.

Complete bibliographic information about a paper by Kostoff et al. (2000) according to the SCI Expanded

```
FN Thomson Reuters Web of Science™
VR 1.0
PT J
AU Wu, M
   Bogaert, J
   D'hooge, J
   Sipido, K
   Maes, F
   Dymarkowski, S
   Rademakers, FE
   Claus, P
AF Wu, Ming
   Bogaert, Jan
   D'hooge, Jan
   Sipido, Karin
   Maes, Frederik
   Dymarkowski, Steven
   Rademakers, Frank E.
   Claus, Piet
TI Closed-chest animal model of chronic coronary artery stenosis.
   Assessment with magnetic resonance imaging
SO INTERNATIONAL JOURNAL OF CARDIOVASCULAR IMAGING
LA English
DT Article
DE Coronary artery stenosis; Ischemia; Myocardial infarction; Magnetic
   resonance imaging; Animal model
ID CARDIAC TROPONIN-I; MYOCARDIAL-INFARCTION; THERAPEUTIC ANGIOGENESIS;
   HEART-FAILURE; ISCHEMIA; INJURY; REPERFUSION; DISEASE; SWINE; DOG
AB To evaluate the consequences of chronic non-occlusive coronary artery (CA) stenosis on myocardial function, perfusion and viability, we developed a closed-chest, closed-pericardium pig model, using
magnetic resonance imaging (MRI) as quantitative imaging tool. Pigs underwent a percutaneous copper-coated stent implantation in the left circumflex CA (n = 19) or sham operation (n = 5). To evaluate
the occurrence of myocardial infarction, cardiac troponin I (cTnI) levels were repetitively measured. At week 6, CA stenosis severity was quantified with angiography and cine, first-pass and contrast-
enhanced MRI were performed to evaluate cardiac function, perfusion and viability. In the stenting group, cTnI values significantly increased at day 3 and day 5 (P = 0.01), and normalized at day 12.
At angiography, 13/19 stented pigs had a stenosis > 75%. Mean degree of CA stenosis was 91 +/- A 4%, range 83-98%. At contrast-enhanced MRI, mean infarct size was 7 +/- A 6%, range 0.7-18.4%. Five of
the 6 pigs with stenosis < 75% had no infarction. Stented pigs showed significantly higher Left-ventricular volumes and normalized mass (P < 0.05), and lower ejection fraction (P = 0.03) than the sham
pigs. Both wall thickening and myocardial perfusion were significantly lower in animals with at least one segment > 50% infarct (23 +/- A 8%; 0.05 +/- A 0.01 a.u./s) and animals with only < 50%
infarct segments (29% +/- A 12%; 0.07 +/- A 0.02 a.u./s), than sham pigs (52 +/- A 6%; 0.10 +/- A 0.03 a.u./s) (P < 0.001; P < 0.05). This minimally-invasive animal model of chronic, non-occlusive CA
stenosis, presenting a mixture of perfusion and functional impairment and a variable degree of myocardial necrosis, can be used as substitute to study chronic myocardial hypoperfusion.
C1 [Claus, Piet] Univ Hosp Leuven, Med Imaging Res Ctr, B-3000 Louvain, Belgium.
   [Wu, Ming; D'hooge, Jan; Sipido, Karin; Rademakers, Frank E.; Claus, Piet] Catholic Univ Louvain, Dept Cardiovasc Dis, B-3000 Louvain, Belgium.
   [Bogaert, Jan; Dymarkowski, Steven] Catholic Univ Louvain, Dept Radiol, B-3000 Louvain, Belgium.
   [Maes, Frederik] Catholic Univ Louvain, Dept Elect Engn ESAT PSI, B-3000 Louvain, Belgium.
RP Claus, P (reprint author), Univ Hosp Leuven, Med Imaging Res Ctr, Campus Gasthuisberg,Herestr 49, B-3000 Louvain, Belgium.
EM piet.claus@med.kuleuven.be
RI bogaert, jan/E-6181-2012; Claus, Piet/C-8529-2013; Maes,
   Frederik/I-7572-2013
OI Maes, Frederik/0000-0003-0027-1479
FU Geconcerteerde OnderzoeksActie (GOA) project (K. U. Leuven, Leuven,Belgium); Flanders Research Foundation (FWO-Vlaanderen, Belgium) [G. 0438.06, G. 0613.09]
FX This work was supported by the Geconcerteerde OnderzoeksActie (GOA) project (K. U. Leuven, Leuven, Belgium) and two research grants (G. 0438.06 and G. 0613.09) from the Flanders Research Foundation
   (FWO-Vlaanderen, Belgium). We thank Mr. Pascal Hamaekers for technical assistance.
```

Source: Clarivate Analytics Web of Science Core Collection

Complete bibliographic information about a paper by Kostoff et al. (2000) according to the SCI Expanded (contd.)

```
CR BOLUKOGLU H, 1992, AM J PHYSIOL, V263, pH20
   Cerqueira MD, 2002, CIRCULATION, V105, P539, DOI 10.1161/hc0402.102975
   Fuchs S, 2001, CORONARY ARTERY DIS, V12, P173, DOI 10.1097/00019501-200105000-00003
   Heinzel FR, 2008, CIRC RES, V102, P338, DOI 10.1161/CIRCRESAHA.107.160085
   Hughes GC, 2003, J APPL PHYSIOL, V94, P1689, DOI 10.1152/japplphysiol.00465.2002
   Kim RJ, 1999, CIRCULATION, V100, P1992
   Klocke R, 2007, CARDIOVASC RES, V74, P29, DOI 10.1016/j.cardiores.2006.11.026
   KLONER RA, 1980, CARDIOVASC RES, V14, P371, DOI 10.1093/cvr/14.7.371
   Leonardi F, 2008, RES VET SCI, V85, P141, DOI 10.1016/j.rvsc.2007.09.010
   Li RK, 2000, J THORAC CARDIOV SUR, V119, P62, DOI 10.1016/S0022-5223(00)70218-2
   PAGANI M, 1978, CIRC RES, V43, P83
   Radke PW, 2006, ENDOTHELIUM-J ENDOTH, V13, P25, DOI 10.1080/10623320600660128
   Reffelmann T, 2004, CORONARY ARTERY DIS, V15, P7, DOI 10.1097/01.mca.0000105482.63241.5f
   Ricchiuti V, 1998, AM J CLIN PATHOL, V110, P241
   ROAN PG, 1981, CIRC RES, V49, P31
   ROTH DM, 1987, AM J PHYSIOL, V253, pH1279
   Sakaguchi G, 2003, ANN THORAC SURG, V75, P1942, DOI 10.1016/S0003-4975(03)00184-X
   Selvanayagam JB, 2005, CIRCULATION, V111, P1027, DOI 10.1161/01.CIR.0000156328.28485.AD
   Song Woohyuk, 2005, J Invasive Cardiol, V17, P452
   Staab ME, 1997, INT J CARDIOL, V58, P31, DOI 10.1016/S0167-5273(96)02844-6
   St Louis JD, 2000, ANN THORAC SURG, V69, P1351, DOI 10.1016/S0003-4975(00)01130-9
   Szilard M, 2000, INT J CARDIOVASC INT, V3, P111
   Terp K, 1999, SCAND CARDIOVASC J, V33, P265
   Weidemann F, 2003, CIRCULATION, V107, P883, DOI 10.1161/01.CIR.0000050146.66577.4B
   Yarbrough WM, 2003, ANN THORAC SURG, V76, P2054, DOI 10.1016/S0003-4975(03)01059-2
NR 25
TC 5
Z9 5
PU SPRINGER
PI DORDRECHT
PA VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS
SN 1569-5794
J9 INT J CARDIOVAS IMAG
JI Int. J. Cardiovasc. Imaging
PD MAR
PY 2010
VL 26
IS 3
BP 299
EP 308
DI 10.1007/s10554-009-9551-1
PG 10
WC Cardiac & Cardiovascular Systems; Radiology, Nuclear Medicine & Medical
   Imaging
SC Cardiovascular System & Cardiology; Radiology, Nuclear Medicine &
   Medical Imaging
GA 575PZ
UT WOS:000276080900007
PM 20012206
ER
```

Source: Clarivate Analytics Web of Science Core Collection

Data cleaning

The bibliometric use of bibliographic data requires careful cleaning of data, notably

- Author names
- Addresses with country/region, institutional information
- Journals, proceedings, books
- Reference items (for various purposes)

Data cleaning

The bibliometric use of bibliographic data requires careful cleaning of data, notably

- Author names
- Addresses with country/region, institutional information
- Journals, proceedings, books
- Reference items (for various purposes)

Furthermore, citation analysis requires redefinition of reference and source items

- reference

Data cleaning

The bibliometric use of bibliographic data requires careful cleaning of data, notably

- Author names
- Addresses with country/region, institutional information
- Journals, proceedings, books
- Reference items (for various purposes)

Furthermore, citation analysis requires redefinition of reference and source items

- reference ⇒ cited item (= new source item)

Data cleaning

The bibliometric use of bibliographic data requires careful cleaning of data, notably

- Author names
- Addresses with country/region, institutional information
- Journals, proceedings, books
- Reference items (for various purposes)

Furthermore, citation analysis requires redefinition of reference and source items

- reference ⇒ cited item (= new source item)
- original source item

Data cleaning

The bibliometric use of bibliographic data requires careful cleaning of data, notably

- Author names
- Addresses with country/region, institutional information
- Journals, proceedings, books
- Reference items (for various purposes)

Furthermore, citation analysis requires redefinition of reference and source items

- reference ⇒ cited item (= new source item)
- original source item ⇒ citing item (= citation)

The use of multiple data sources

The efficiency of bibliometric work can be enhanced by combining different data sources.

- ☞ Different sources follow different standards, some do not follow any standard.

The combination of a citation database with a specialised database with an excellent hierarchical subject classification scheme might improve coverage and precision of the results.

The use of multiple data sources

The efficiency of bibliometric work can be enhanced by combining different data sources.

- ☞ Different sources follow different standards, some do not follow any standard.

The combination of a citation database with a specialised database with an excellent hierarchical subject classification scheme might improve coverage and precision of the results.

- ☛ Appropriate data cleaning is indispensable for matching data from different data sources.

Elsevier's SCOPUS database

Elsevier SCOPUS is a second large multidisciplinary abstract and citations database for scholarly journals. It is provided by Elsevier since 2004.

The thematic coverage is similar to that of the WoS. The number of journals covered by SCOPUS is, however, larger than that of the WoS. Besides journals also proceedings and book series are covered.

Similarly to the previously introduced WoS, SCOPUS licenses to off-line custom data and is fully featured for bibliometric use.

SCOPUS provides a large range of tools closely connected to bibliometrics: analytics (including journal analyser), citation statistics including elementary measures (h-graph etc.).

☞ Users should notice that Scopus and Web of Science are differently structured. This is due to historical reasons, the underlying data sources and processing as well to their policies.

Google Scholar

Google Scholar provided by *Google* enables searches for scholarly publications including "*peer-reviewed papers, theses, books, preprints, abstracts and technical reports from broad areas of research*".
The service is based on the Google search engine.

Google Scholar

Google Scholar provided by *Google* enables searches for scholarly publications including "*peer-reviewed papers, theses, books, preprints, abstracts and technical reports from broad areas of research*".
The service is based on the Google search engine.

Google Scholar does not meet all requirements of advanced bibliometric studies.

- strictly speaking not a database, uses web links and provides "secondary" information
- proper documentation and crucial information about coverage is missing
- no acceptable disambiguation for publications and citations
- citations are leaving the "publication universe"
- sources partially questionable

Google Scholar

- assignment to addresses is not always possible
- no native subject classification
- access to reference sets for large-scale analysis not possible
- considerable changes of sources and citations within shorter time periods due to

Google Scholar

- assignment to addresses is not always possible
- no native subject classification
- access to reference sets for large-scale analysis not possible
- considerable changes of sources and citations within shorter time periods due to

$\Rightarrow$ Google Scholar is a useful retrieval tool for author and topic searches with interesting link features but limited search fields. The database is, due to its setup and structure, not suited for systematic, comparative and large-scale bibliometric application.

End of Part I

Thank you very much for your attention!

Professor Dr. Wolfgang Glänzel

Wolfgang.Glanzel@kuleuven.be
feb.kuleuven.be/Wolfgang.Glänzel

Researcher ID: A-6280-2008
ORCID: 0000-0001-7529-5198
Scopus Author ID: 7003697821
MR Author ID: 74130