Semantics and Bibliometrics in Educational Research: State-of-the-art report

Åström, Fredrik

2010

EUROPEAN EDUCATIONAL RESEARCH
Quality Indicators

A project funded under the Socio-Economic Sciences and Humanities Theme (SSH)

SEVENTH FRAMEWORK PROGRAMME

Semantics and Bibliometrics in Educational Research

Fredrik Åström
Lund University Libraries
Fredrik.astrom@lub.lu.se

Introduction

At the beginning of the EERQI project, one of the principal aims was to investigate the possibility to identify keywords reflecting research quality by means of automatic semantic analyses. In relation to another central purpose of the project, the development of a search engine, attempts has been made, using semantic analyses for automatic identification of descriptor keywords of documents in the EERQI Content Base (CB). Both these lines of inquiry has proven to be problematic; which – in combination with additional problems related to extracting cited references from the CB documents – has left originally intended analyses on relations between keyword and reference/citation structures hard to pursue.

The purpose of this report is to answer the following questions:

- What structures we can identify by applying bibliometric methodology to semantic information related to the BD documents?
- How can these structures – or lack thereof – help us explain the problems related to the attempts at indentifying semantic quality indicators as well as document descriptors?

Thus, this report does not primarily apply bibliometric methods for analyzing research productivity or the impact of research through publication or citation analysis – the line of analysis that is most often associated with bibliometrics – but is rather relating to the tradition of using bibliometric methodology for identifying structures by quantitative analyses of texts and text representations in a wider sense. However, in addition to the abovementioned questions, an analysis of how descriptor type information can give us some information of the visibility of European educational research is reported.

Three sets of frequency and co-occurrence analyses were performed on three different data sets; all data related to the EERQI CB, albeit from different perspectives. The first set of analyses was done on automatically extracted descriptors from 35 articles in ten CB journals, a work performed by Xerox XRCE. For the second and third sets, a different approach was required. The starting point was the 100 articles forming the basis for the peer review exercise as well as the bibliometric analyses performed by the Berlin School of Library and Information Science at Humboldt University (HU-IBI), from which information on publishing journals as well as article authors was collected. Based on this information, the *Educational Re-*

sources Information Centre (ERIC) and Web of Science databases were searched to collect additional data to analyze.

XRCE

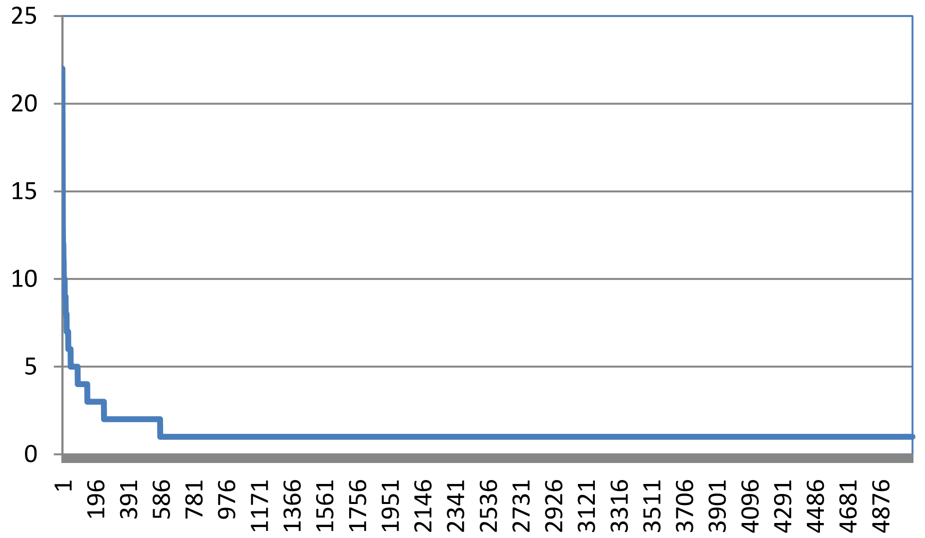
The aim of the first set of analyses is to analyze structures in descriptors automatically identified in the CB documents, and is thus based on the work of Xerox XRCE where they used noun phrase extraction techniques to identify descriptor type keywords in 35 articles from ten journals in the CB (Table 1).

Table 1. Source journals for XRCE noun phrase extraction analysis

British Journal of Sociology of Education
Child Development Perspectives
Comparative Education
Educational and Psychological Measurement
Educational Psychologist
Educational Researcher
Educational Theory
Gender and Education
History of Education
Sociology of Education

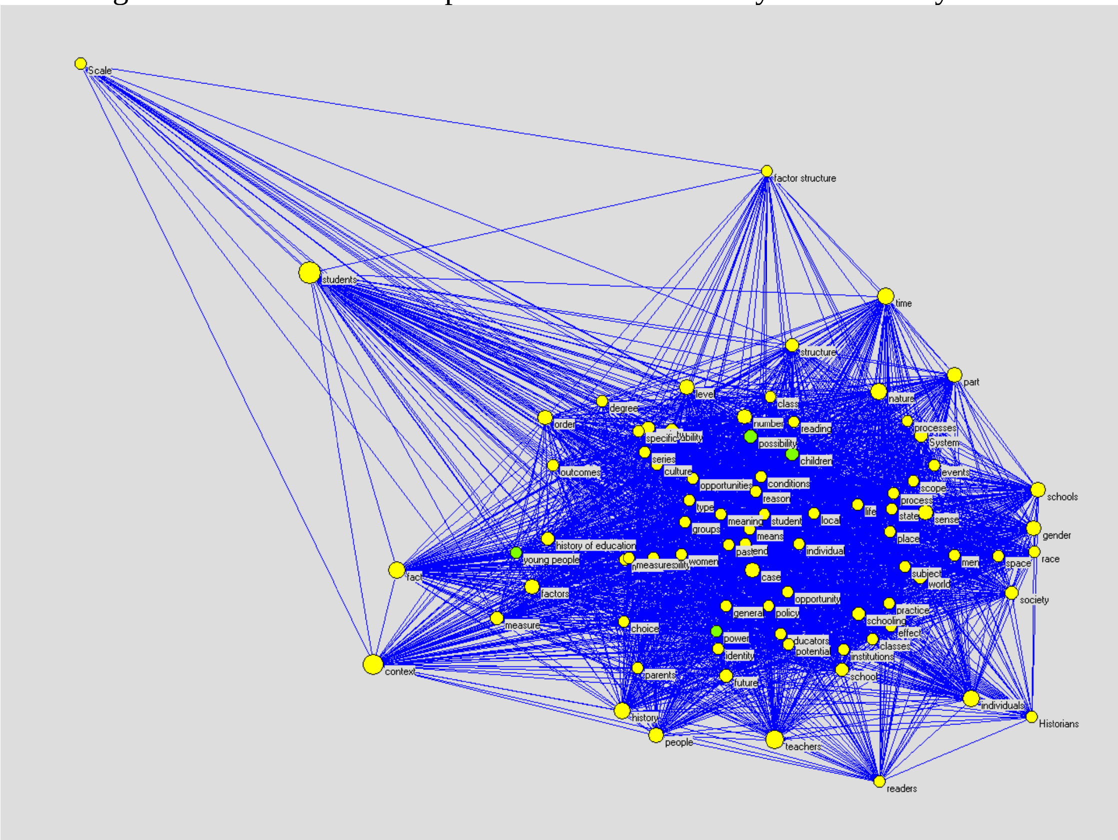
From these 35 articles, a total of 6,324 nouns and noun phrases were extracted; with a number of 5,069 unique keywords. As the 1.25 total to unique keyword ratio suggest – and as can be exemplified by the frequency/keyword distribution (Figure 1) – there are few keywords occurring more than once; and even the most frequently occurring keyword is only present 22 times in the data set.

Figure 1. XRCE keyword frequency distribution



To investigate the relation between the keywords; and to see if they could be used to say something about the intellectual structure of the content of the articles: a co-occurrence analysis of the keywords was made. Based on keywords occurring more than five times and to what extent these keywords occur together in the documents – the more documents they co-occur in, the stronger the relation between the keywords is – a map of keyword relations can be constructed (Figure 2). This method was suggested by e.g. Whittaker (1989) to identify different areas of interest in wider research fields; but when applied on this material, the lack of noticeable structures is evident.

Figure 2. Co-occurrence map of XRCE automatically extracted keywords



The lack of identifiable structures is largely depending on the low occurrence frequencies of even often occurring keywords; and although there are many connections between the words, the connections are weak and contain little meaning. In addition to the issue of low frequencies and co-occurrence strength, the keywords identified are also relatively non-specific, saying little about the content of the research articles: especially in terms of research foci such as theoretical or methodological orientation, research perspective and so on.

The results of the co-occurrence analysis build on the pair-wise relations between the keywords, but these were weak and provided little in legible structures. However, when raising the demands on the connection strength between keywords by using a clustering routine suggested by Persson (1994), the keywords and their inter-relation started to make more sense (Table 2). The principle of the clustering routine is basically to join pairs with one common unit: e.g. pair A – B and pair B – C will form a cluster, whereas pairs A – B and C – D will not. The results are not unambiguous, there are overlaps between e.g. cluster five (‘female’) and cluster nine (‘women’); and it should be kept in mind that the analyses are based on a small sample of only 35 articles from ten journals, where the structure identified might be more of a reflection of the content of individual articles/journals, rather than a more general structure of different research orientations in educational research. However: the stricter demands on how to join individual keywords did bring similar issues together; and it is also interesting to note how the extraction of keywords also ‘brought along’ author names that the articles are referring to; and that they do so in a way that seem to correlate to the keywords in a good way.

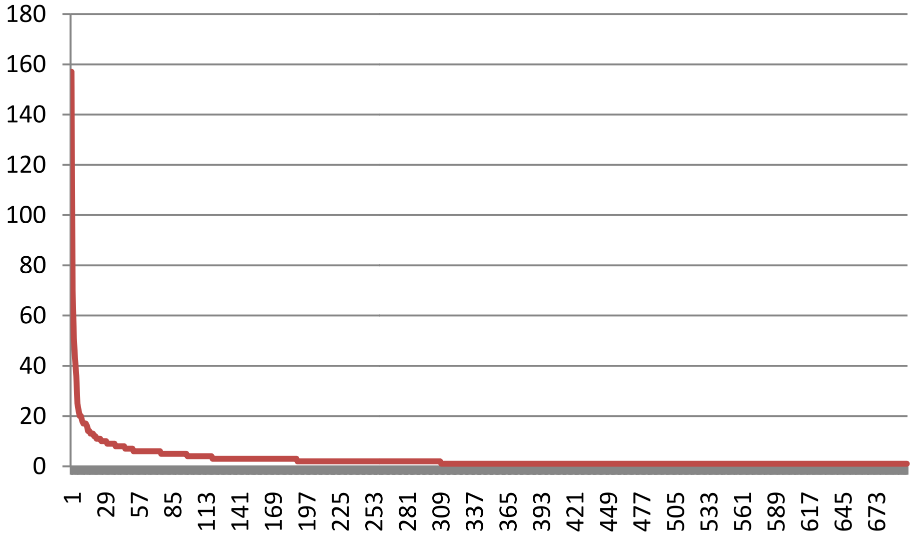
Table 2. Clustering of XRCE keywords (selection)

Cluster	Keywords
1	students prior knowledge; opportunity factors; Disabilities; System; history; practice; race-ethnicity; factors; factor structure
2	Durkheim; approach to pedagogy; generalised other; identity; moral identity; organisation; pedagogic communication; phenomena; sociology
3	Ethnic Studies programs; pattern; economics; ethnicity; existence; more selective institutions; personal physicians clearance; physical ability; Ramirez
4	social provisions; early adolescents; multivariate normal distribution; stability; scale; Davies; adolescents appraisals of interparental conflict; cross-sectional; early and late adolescent samples
5	Female; average salary; correlation; negative; penetration of women; percentage female; positive effect; queuing perspective states; range
6	Problematic; educational settings; modern capitalist society; problematize; proletarian souls; public consciousness; racial inequalities in educational opportunities; revolutionary educators; social groupings
7	oral testimony; oral history; narrative identities; narrative identity; Paul Thompson; pristine subjectivity; private memory; profoundest suspicion; qualitatively different testimony
8	Schieder; Moglichkeiten; physics; practice of comparison; R Wittram; radius; rational thinking; real course; necessity
9	Men; women; AMAS-C; majority; Physiological Anxiety; plausible reliability estimates; postrotation variance; psychometric soundness; college students
10	other buildings; college; library; Centres; physical or social; physical site; proactive stance; professions; programme

ERIC

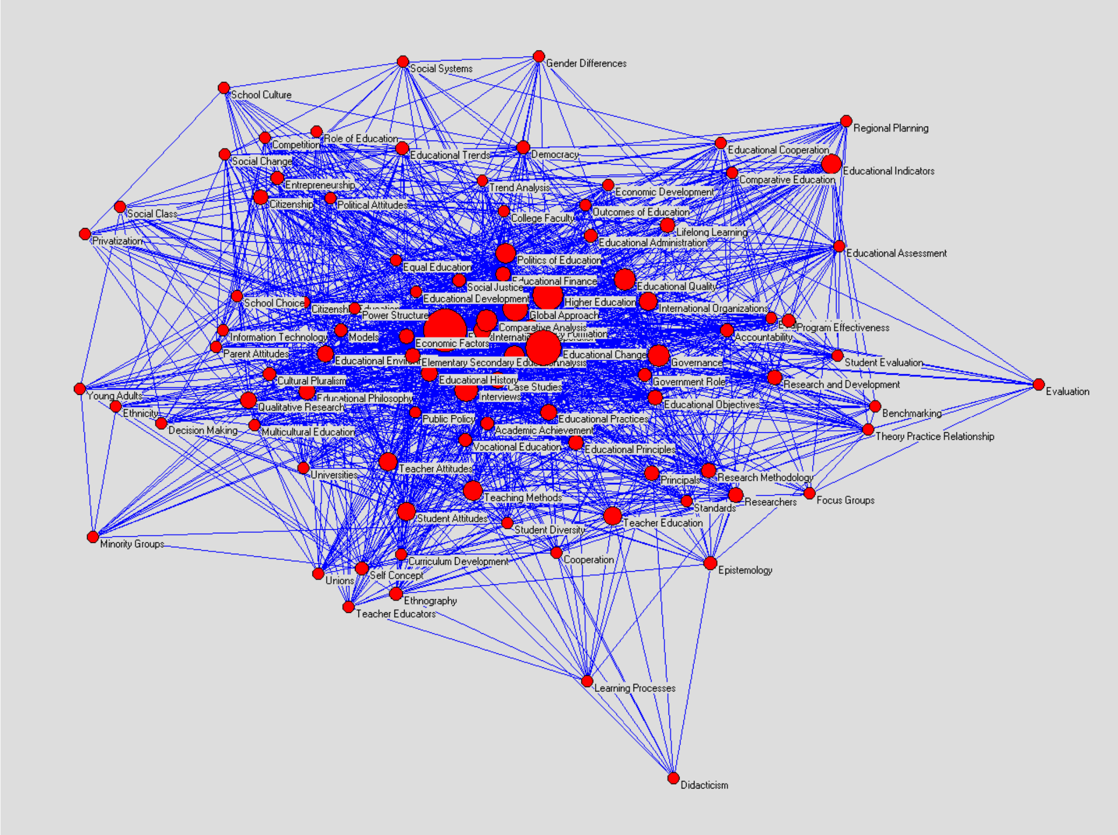
To be able to compare the keywords extracted by XRCE with a controlled set of descriptors, analyses of EERQI CB documents was also done using data from the ERIC database. However, since the options for downloading data from the ERIC web interface is limited, only metadata for research articles from *European Education Research Journal* (EERJ), published 2000-2010, were downloaded (ERIC search: Source – European Educational Research Journal; Publication date – 2000-2010; Publication type – Journal Articles). The download contained 188 documents described with 699 unique descriptors used 2,065 times (10.98 descriptors/document). As opposed to the automatically extracted keywords, there are substantially fewer keywords per document; at the same time as the number of documents are higher, resulting in a larger concentration of the keywords and also; a higher amount of keywords occurring several times as well as the often occurring keywords shows much higher frequencies than in the XRCE material (Figure 3).

Figure 3. Frequency distribution of ERIC Descriptors



As with the XRCE keywords, the relations between the keywords were investigated by a co-occurrence analysis; and as with the XRCE co-occurrence map, there are few identifiable structures: the majority of the keywords gathers in the centre of the map (Figure 4). However, unlike the XRCE keywords, the results of the cluster analysis described above, did not yield any meaningful results either: all keywords were gathered in one big cluster.

Figure 4: Co-occurrence of ERIC descriptors



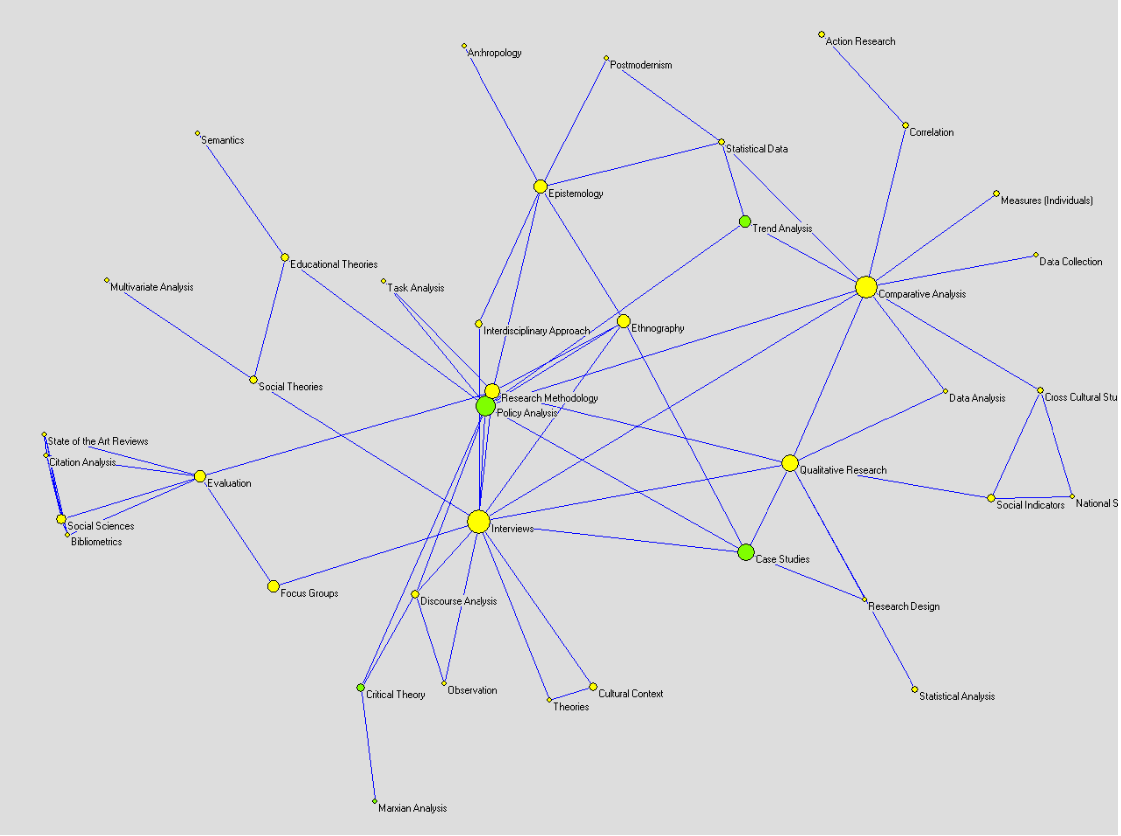
Since the keywords retrieved from the ERIC descriptor field is more specific than the XRCE keywords, one would expect the ERIC descriptors to respond better to the quantitative analyses. However, an explanation can be found when looking into the nature of the descriptors (Table 3).

Table 3: ERIC descriptor 'facets'

Method	Educ. Level	'Perspective'	'Issues'
Comparative Analysis	Higher Education	Educational Policy	Ethnicity
Case Studies	Secondary Schools	Educational Philosophy	Democracy
Content Analysis	Preschool Education	Educational History	Social Justice
Discourse Analysis	Postdoctoral Education	Educational Technology	Theory-Practice Rel.

The descriptors make no distinction between keywords describing methodological, theoretical or empirical issues, issues that can be described as facets of the document description. As e.g. the descriptor 'discourse analysis' can be combined with both 'preschool' and 'postdoctoral' education, as well as with 'policy' or 'history' perspectives and 'ethnicity' and 'democracy' issues; on an aggregated level, e.g. methodological distinctions will be eliminated by all the combinations possible in the other facets. Thus, if we make a co-occurrence analysis restricted within one facet, we start seeing more legible structures in the map (Figure 5).

Figure 5. Facet limited co-occurrence of ERIC descriptors



ISI

So far, this report has primarily been focused on investigating problems relating to identifying descriptors for classification purposes, as well as keywords for use as quality indicators. This last set of analyses, however, takes on another approach; investigating how we can use keyword type information for investigating the visibility of European educational research. Thus, in this section, we do not study documents within the EERQI CB, but documents citing CB articles. To find these, the 'Cited Reference Search' (CRS) option in Web of Science (WoS) was used; and the starting point for data selection was the 100 EERQI articles analyzed by HU-IBI. In CRS, full journal names cannot be used for searching; and preferred journal title abbreviations are only listed for journals indexed in WoS, while none of the 100 EERQI articles are indexed in WoS. Also, very few of the 100 articles are cited in WoS-indexed journals; however, by searching cited EERQI authors in CRS, we can retrieve information about journal title abbreviations for those articles that actually are cited. This way, it was possible to find data for analyzing the visibility of European educational research (Table 4).

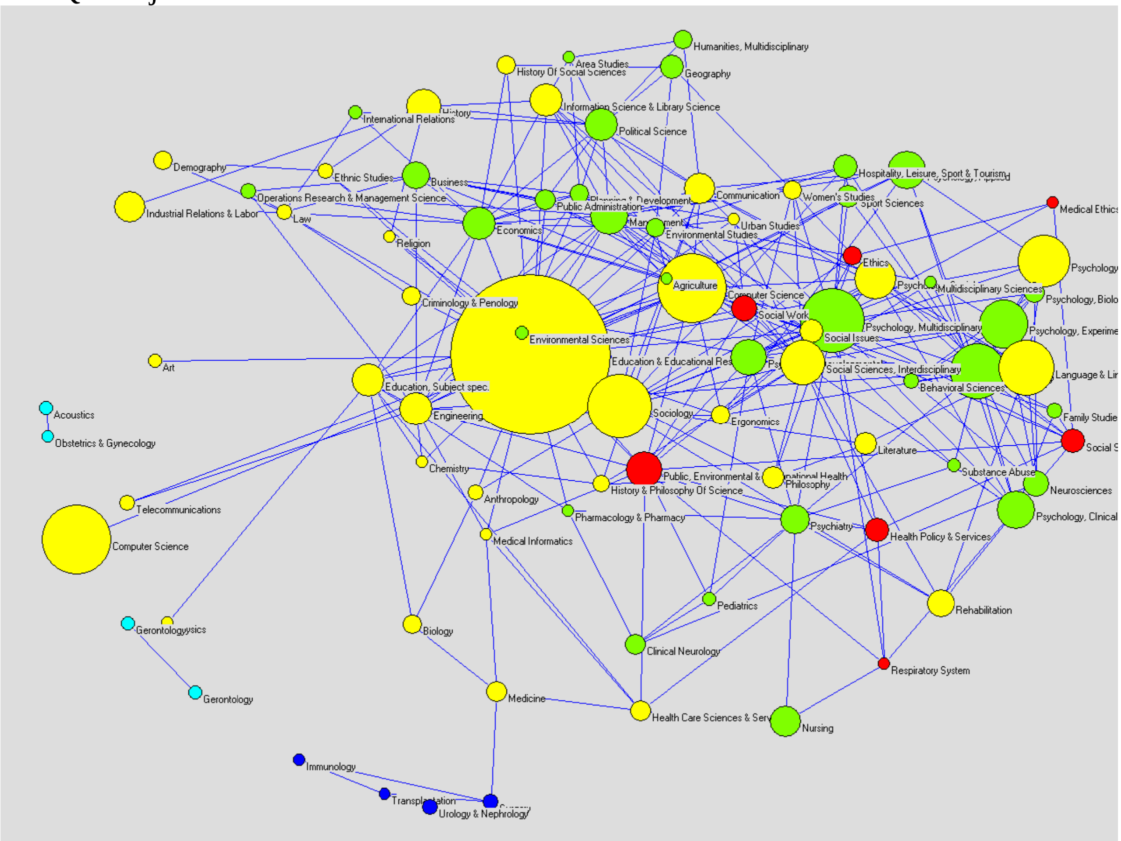
Table 4. EERQI CB journals cited in ISI-indexed journals

Journals	Cited articles	Citing articles
Z DIDAKTIK NATURWISS	49	67
ERZIEHUNGSWISSENSCHA	480	473
FORUM QUALITATIVE SO	459	553
CONTEMPORARY ISSUES	276	475
EUROPEAN ED RES J	120	180
E LEARNING	107	163
POLICY FUTURES ED	97	132
REV FRANCAISE PEDAGO	366	400
RES COMP INT ED	18	19

As might be expected, there was some overlap in citations between the different journals; however, the overlap was relatively small: out of the total of 2,444 articles citing EERQI CB journals, only 18 of those cite more than one of the RRQI journals. To investigate in what fields EERQI journals are cited – thus investigating in what research fields those articles are visible – analysis were made on the subject categories (SC) from Web of Science: Journal Citation Reports (JCR), categories that are used for classifying the journals indexed in the WoS databases.

When analyzing subject categories used for classifying journals with articles citing EERQI journals: 160 subject categories were found. Not surprisingly, ‘Education & Educational Research’ is by far the most frequently used category; and neither surprising is the high frequencies of ‘Computer Science’ and various psychology oriented categories, both of which are research fields with significant overlaps to educational research. Aside from these, other highly represented subject categories are e.g. ‘Sociology’, ‘Social Science, Interdisciplinary’ and ‘Language & Linguistics’. To see how the subject categories relate to each other, the same co-occurrence analysis was made again (Figure 6). It should be noted that the relations between the subject categories are based on how they are used together for classifying journals, i.e. drawing on journals being categorized as both being e.g. a computer science journal and an educational research journal. Therefore, to be able to say anything about the visibility of educational research outside its own realm (as defined by the journals included in the JCR subject categories), we cannot look at the frequencies of other subject categories without taking the double classification into account.

Figure 6. Co-occurrence of JCR subject categories for journals containing articles citing EERQI CB journals



To be able to say anything about the visibility of educational research outside its own confines, we need to limit the analysis to the journals categorized as e.g. ‘sociology’ journals that are not also classified as ‘educational research’ or ‘psychology, educational’. After the elimination of these, to get a fair representation of the frequency distribution between the subject categories, we also need to fractionalize the counts for other double categorizations. Thus, if a journal is categorized as both ‘sociology’ and ‘computer science’: each category gets a frequency of 0.5 (Table 5).

Table 5. Fractionalized frequencies of JCR subject categories for journals citing EERQI documents, ‘Educational Research’ and ‘Psychology, Educational’ excluded

Frequency	Subject category	Frequency	Subject category
120	Psychology, Multidisciplinary	25	Economics
119	Sociology	22	Psychiatry
63	Computer Science	22	Communication
54	Psychology, Experimental	22	Information Science & Library Science
53	Social Sciences, Interdisciplinary	21	Industrial Relations & Labor
53	Psychology	19	Social Work
50	Psychology, Social	18	Geography
47	Language & Linguistics	19	Literature
39	Psychology, Clinical	16	Business
32	Nursing	15	Rehabilitation
31	Management	15	Medicine
31	Psychology, Developmental	15	Philosophy
29	Political Science	14	Health Policy & Services
28	History	13	Engineering
26	Psychology, Applied	13	Demography
25	Public, Environmental & Occupational Health	12	Criminology & Penology

To further investigate the visibility of European educational research outside the research field itself; the distribution of frequencies over time were analyzed (Figure 7). Although the figures are small on subject code level, an increase in frequencies can be seen in the 2000s, with e.g. 'history', 'communication' and 'sociology' presenting relatively high figures.

Figure 7a. Annual frequency distribution of JCR subject categories for ISI journals citing EERQI journals; educational research, computer science and psychology excluded

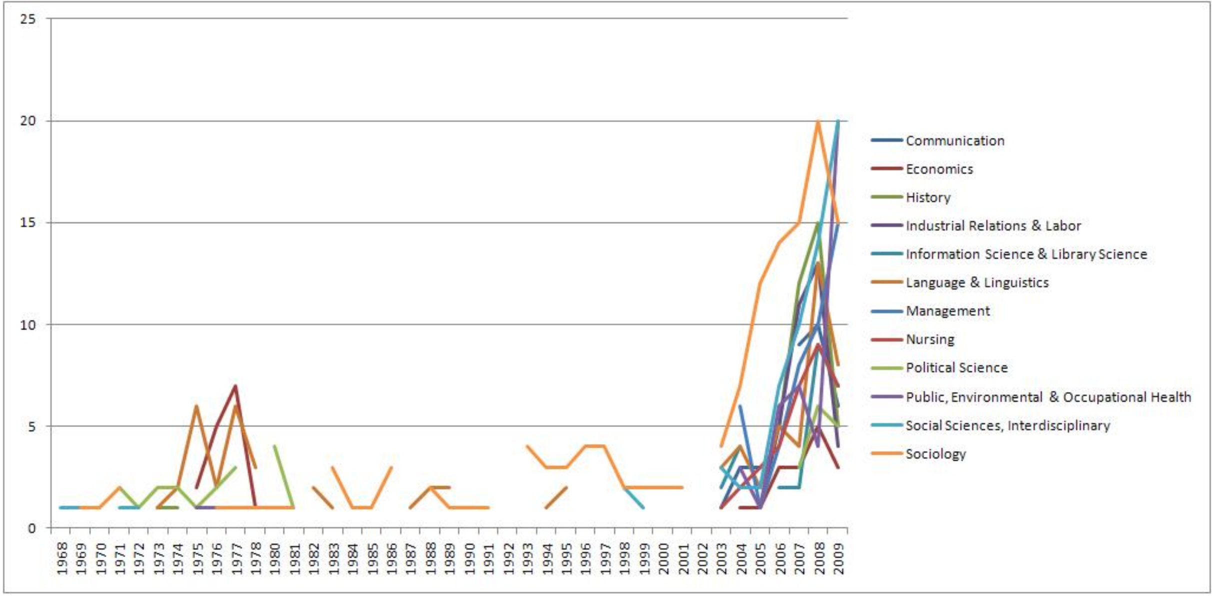
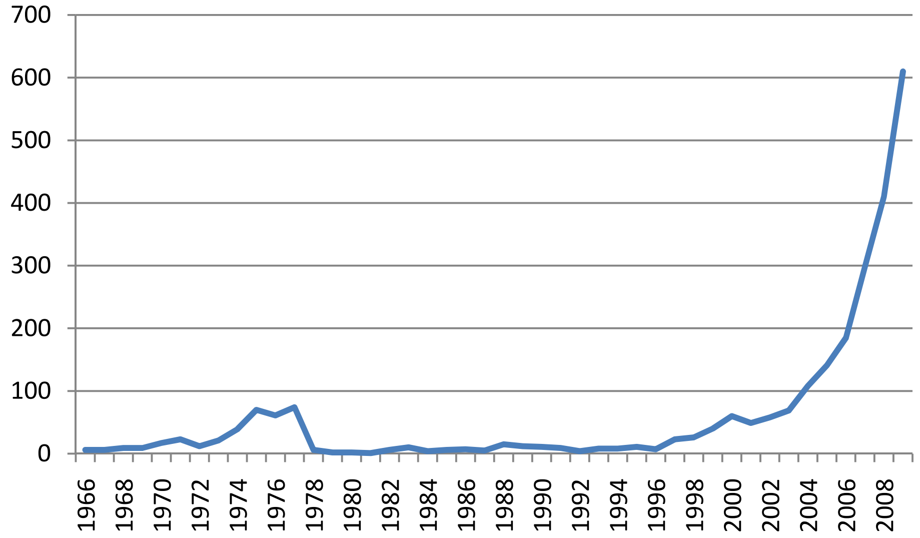


Figure 7b. Yearly distribution of ISI articles citing EERQI journals (for comparison purposes)



Discussion

The main issue in this report is the problems related to different applications of automatic keyword extraction. A main reason for this seems to be how very few keywords occurs more than a few times in any document set analyzed here: the same phenomenon was also seen in analyses not reported here, such as an analysis on the WoS material with descriptors for the documents citing EERQI articles, as well as poor results in analyses grouping the documents based on shared properties. Another problem related to the keyword analyses is how one 'list' of keywords brings together different levels of descriptions, as discussed in relation to the ERIC descriptors. This makes it hard to perform any quantitative or automatic analyses on the keywords, regardless if we are dealing with more or less specific and specialized keywords.

The background of this can be discussed on different levels. On one level, there is the issue of terminology, where e.g. Richard Whitley (2000) describes fields characterized as ‘fragmented adhocracies’ as fields with, among other traits, a terminology that is rather characterized by the use of ‘everyday’ language and ‘layman’ terms rather than a high level of specialization; and also, a relatively low level of consensus, both in terms of terminology as well as e.g. on matters of methodology and work techniques. On another level, there is the issue of the multidisciplinary nature of educational research, in combination with the close relationship to the field of professional practice and different kinds of educational institutions. This presents us with a field of a heterogeneous nature, where not only can certain phenomena or processes be investigated from different points of view, but there is also processes and phenomena that demands widely different approaches for analysis; a trait that educational research shares with e.g. library and information science (LIS) and that presents both risks and opportunities (Nolin & Åström, 2010). One important aspect of this is of course also the variations in e.g. social, cultural and professional contexts that educational research is both studying and working within, with differences e.g. between educational systems in different countries as well as the systems for different levels of education.

There are results in this report that can be cause for further inquiry. The cluster analysis of the XRCE keywords worked quite well, but it needs to be investigated whether the results primarily comes out of the small empirical material. Another path to inquire further into is whether a combination of the ERIC descriptors and a co-citation analysis would help to distinguish better between different areas within educational research.

Finally, a few words on the analysis based on articles in WoS citing EERQI journals. It is interesting to note how the visibility of educational research seems to be increasing, not only in terms of European educational research becoming more visible in general, but also, the visibility in relation to other fields of research. One such example is the relation between educational research and LIS, where LIS has started to show an increasing interest in educational/pedagogical issues, for instance in relation to inquiries into ‘information literacy’; and that can also be exemplified by mutual research projects such as LinCS (http://www.ipd.gu.se/english/Research/research_programmes/lincs/) at Gothenburg University and the University College in Borås.

References

Nolin, J. & Åström, F. (2010). Turning weakness into strength: Strategies for future LIS. *Journal of Documentation*, 66(1), 7-27.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science,* 45(1), 31-38.

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford: Univ. Press.

Whittaker, J. Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science*, 19(3), 473-496.

Figure 4: Co-occurrence of ERIC descriptors

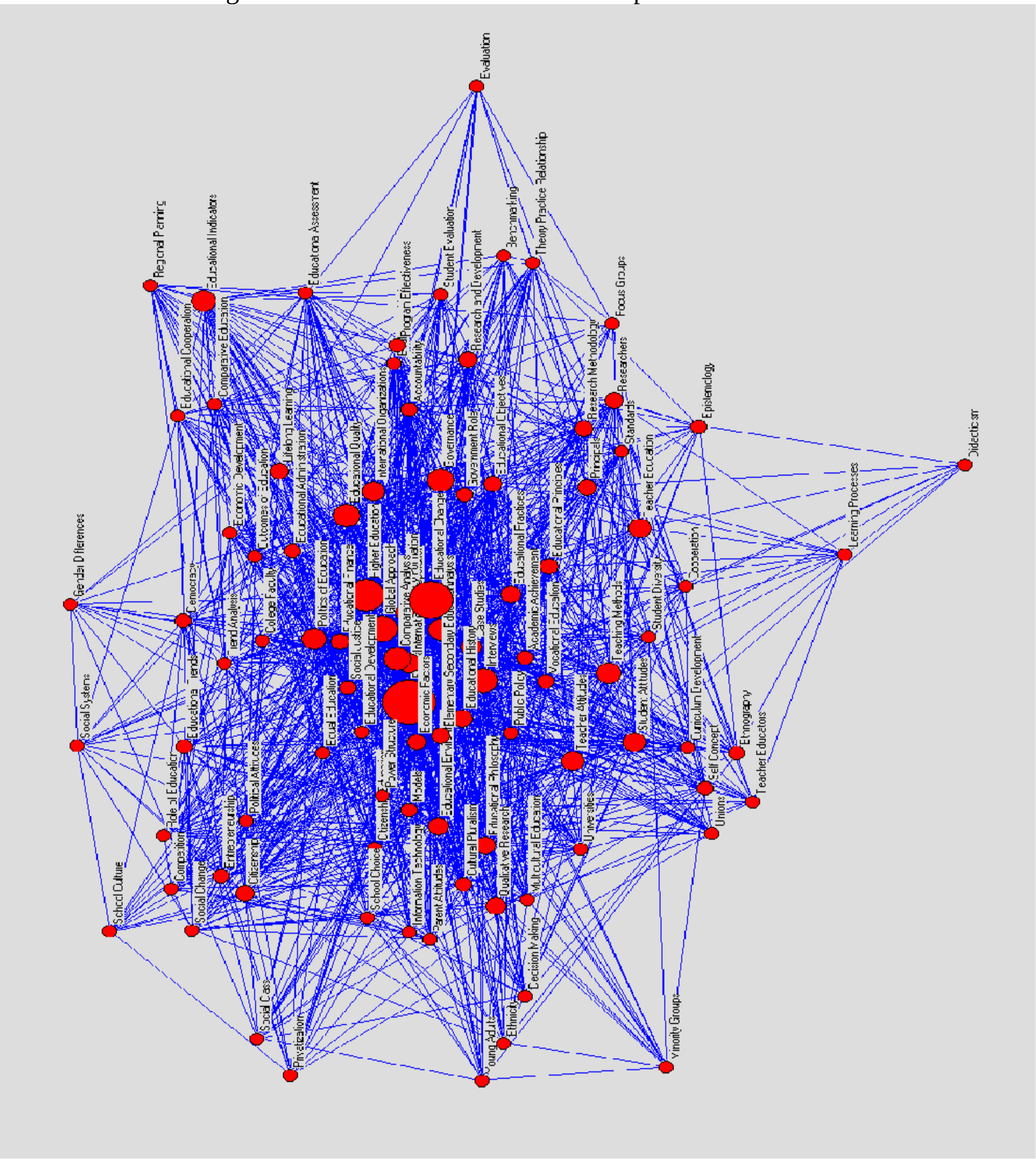


Figure 5. Facet limited co-occurrence of ERIC descriptors

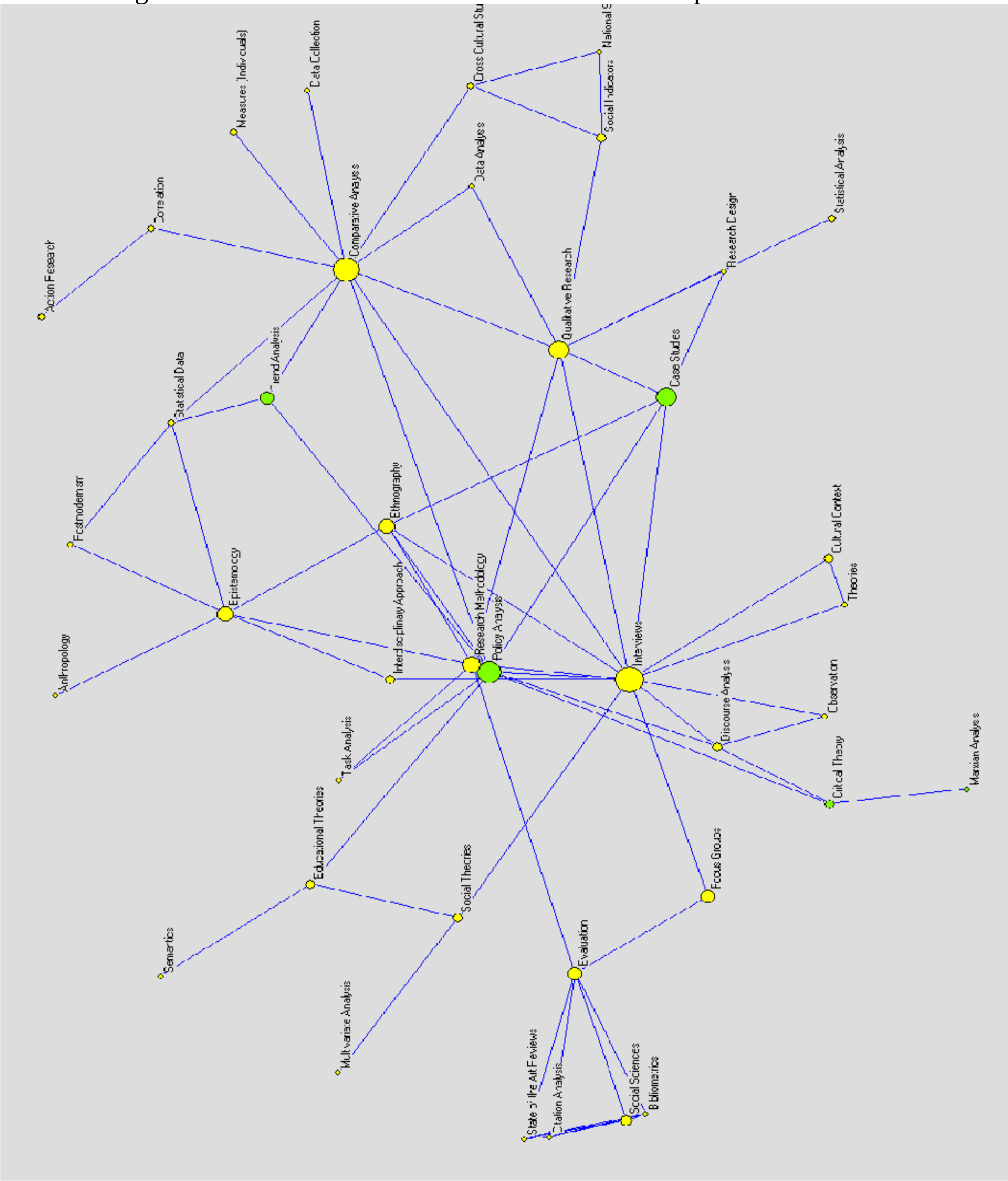


Figure 6. Co-occurrence of JCR subject categories for journals containing articles citing EERQI CB journals

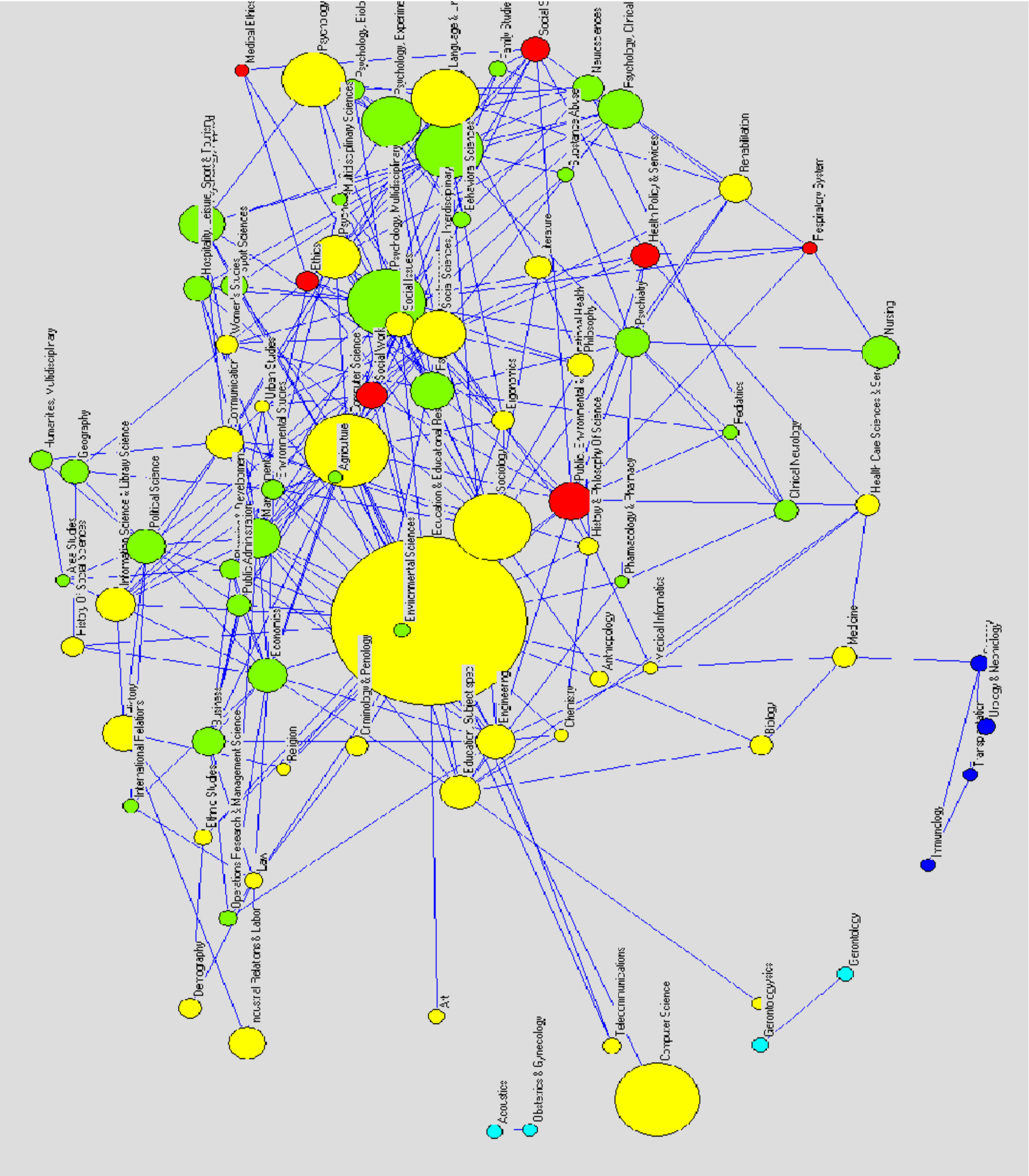
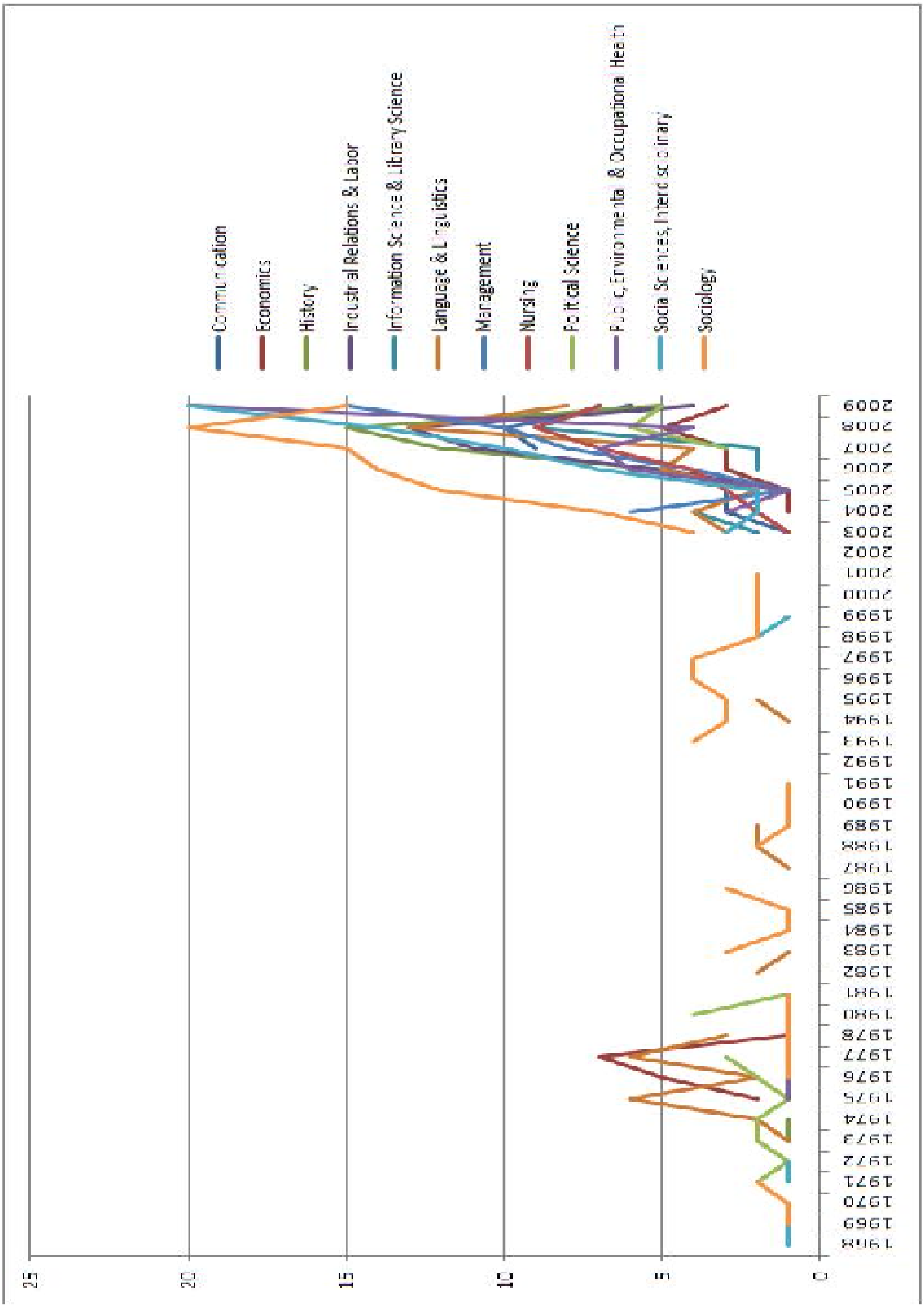


Table 5. Fractionalized frequencies of JCR subject categories for journals citing EERQI documents, ‘Educational Research’ and ‘Psychology, Educational’ excluded

120,247	Psychology, Multidisciplinary	6,332	Health Care Sciences & Services	1,499	Otorhinolaryngology
118,832	Sociology	6,119	Planning & Development	1,476	Ecology
62,878	Computer Science	6	Ethics	1,333	Urban Studies
53,863	Psychology, Experimental	6	Family Studies	1,333	Social Sci, Mathematical Methods
53,448	Social Sciences, Interdisciplinary	6	History & Philosophy Of Science	1,333	Biotech & Applied Microbiology
53,195	Psychology	5,975	Environmental Studies	1,25	Physics
49,915	Psychology, Social	5,75	Sport Sciences	1,25	Oncology
47,326	Language & Linguistics	5,666	Pediatrics	1,166	Crystallography
39,497	Psychology, Clinical	5,332	Surgery	1,166	Geosciences, Multidisciplinary
32,333	Nursing	5	Gerontology	1	Psychology, Psychoanalysis
31,033	Management	5	Ethnic Studies	1	Theater
30,997	Psychology, Developmental	4,666	Substance Abuse	1	Dermatology
29,165	Political Science	4,333	International Relations	1	Metallurgy & Metallurg Eng
28	History	4,332	Behavioral Sciences	1	Microscopy
25,914	Psychology, Applied	4,083	Ergonomics	1	Hematology
24,964	Public, Envir & Occup Health	4	Art	1	Peripheral Vascular Disease
24,617	Economics	3,833	Obstetrics & Gynecology	1	Archaeology
22,498	Psychiatry	3,75	Law	1	Endocrinology & Metabolism
22,163	Communication	3,666	Pharmacology & Pharmacy	1	Education, Subject spec.
21,665	Information Science & Library	3,582	Operations Res & Managem Sci	1	Forestry
20,5	Industrial Relations & Labor	3,5	Nutrition & Dietetics	1	Gastroenterology & Hepatology
18,666	Social Work	3,5	Chemistry	0,916	Water Resources
17,5	Geography	3,5	Biology	0,833	Transportation Sci & Technology
17,5	Literature	3,476	Agriculture	0,833	Transportation
15,724	Business	3,333	Veterinary Sciences	0,75	Statistics & Probability
14,999	Rehabilitation	3,2	Multidisciplinary Sciences	0,666	Meteorology & Atmospheric Sci
14,833	Medicine	3	Religion	0,666	Energy & Fuels
14,5	Philosophy	3	Physiology	0,583	Genetics & Heredity
14,132	Health Policy & Services	2,866	Telecommunications	0,5	Music
12,759	Engineering	2,725	Environmental Sciences	0,5	Medical Laboratory Technology
12,5	Demography	2,666	Medical Informatics	0,5	Marine & Freshwater Biology
11,5	Criminology & Penology	2,666	Area Studies	0,5	Biochemistry & Molecular Biology
11,333	Humanities, Multidisciplinary	2,5	Ophthalmology	0,5	Reproductive Biology
11,331	Neurosciences	2,5	Medical Ethics	0,5	Orthopedics
10,416	Hospitality, Sport & Tourism	2	Folklore	0,333	Biodiversity Conservation
9,666	Public Administration	2	Pathology	0,333	Film, Radio, Television
8,883	Social Sciences, Biomedical	2	Psychology, Educational	0,333	Cardiac & Cardiovascular Systems
8,75	Social Issues	2	Infectious Diseases	0,333	Toxicology
8,5	Urology & Nephrology	2	History Of Social Sciences	0,333	Radiol, Nuclear Med & Med Imag
8,333	Women's Studies	1,832	Immunology	0,333	Mathematics, Interdisc App
8	Anthropology	1,832	Transplantation	0,25	Materials Science, Multidisc
7,416	Psychology, Biological	1,633	Respiratory System	0,25	Optics
7,333	Acoustics	1,5	Dentistry, Oral Surg & Medicine	0,25	Imaging Sci & Photographic Tec
7,333	Clinical Neurology	1,5	Psychology, Mathematical		

Figure 7. Annual frequency distribution of JCR subject categories for ISI journals citing EERQI journals; educational research, computer science and psychology excluded


Communication
Economics
History
Industrial Relations & Labor
Information Science & Library Science
Language & Linguistics
Management
Nursing
Political Science
Public, Environmental & Occupational Health
Social Sciences, Interdisciplinary
Sociology