Data integration in Scandinavia

Gunnar Sivertsen

gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education (NIFU)
P.O. Box 2815 Tøyen, N-0608 Oslo, Norway

Abstract

Recent developments in Scandinavia may be of interest in relation to developing an integrated European research information structure that could provide the basis for an improved knowledge base for research policy. This article describes how reliable bibliographic data in institutional or national research information systems have been developed in Denmark, Finland, Norway and Sweden with performance-based institutional funding models as a driver. It also discusses in more general terms the limitations and potentials of using data from research information systems in bibliometric analysis and in social studies of science.

Keywords

Research information systems; data integration; bibliometrics; performance-based funding; gender; age; productivity; Scandinavia

Introduction

Current trends in the extended use of research information systems at the institutional, national and European level may give an opportunity for the development of an integrated European research information structure that could provide the basis for an improved research policy development in Europe (Bahieu et al., 2014). These trends have also been regarded as opening the possibility of an increased representation and visibility in bibliographic databases of the scholarly publishing in the social sciences and humanities (Hicks & Wang, 2009). The basic idea is that the present development of institutional and national information systems can become the pillars of an integrated international system. Consequently, it has become important to monitor and coordinate the present national developments.

Promising technical development in this area is currently performed by collaborating organizations such as EUROCRIS, CASRAI and ORCID, and by the providers of commercial solutions to research information systems, such as the PURE system by Elsevier and the CONVERIS system by Thomson Reuters. Non-commercial solutions on the national level have been created in several countries as well, e.g. the R&D Information System in the Czech Republic, the Estonian Research Information System, and the CRISTIN system in Norway. All technical developments are, however, also dependent on the policy level and on the interaction between research organizations, their funders, and the expertise in research information, documentation, indicators and evaluation. Another important driver in these developments seems to be the introduction of performance-based funding systems because they require standardized data at the institutional level.

All of the general developments mentioned above have been present in the Scandinavian countries for several years already. Recently, Denmark, Finland, Iceland, Norway and Sweden engaged in formal collaboration on the creation and analysis of comparable data from national current research information systems. These countries are therefore of interest from the

Publisher version: DOI 10.1007/s11192-015-1817-x

perspective of monitoring and coordinating national developments. As we shall see, there are also different lessons to be learnt from different developments in each of the countries.

My background for giving this short introduction to the Scandinavian experiences is that I have been a formal or informal advisor to the developments in all of the countries except Iceland. I also developed the so-called "Norwegian model" (Sivertsen, 2010), which has been implemented in three of the countries, as explained below. I begin this article by giving an overview of the different combinations of performance-based funding and data integration in Denmark, Finland, Norway and Sweden. I describe the experiences so far with similar, but partially different solutions. Then follows a more general discussion of the limitations and potentials of using data from current research information systems for bibliometric analysis. I illustrate the potentials by reporting from a study of age, gender and productivity in complete data covering scientific publishing at Norwegian research institutions.

Performance-based funding and data collection in Scandinavia

The funding of research institutions is partly performance-based in most European countries (Hicks, 2012). Best known is the research assessment exercise in the United Kingdom, recently implemented as the Research Excellence Framework in 2014. The idea of using the results of research evaluation in institutional funding has partly been introduced in Italy as well. The Czech Republic and Sweden are presently considering a similar procedure. The more widespread solution among the smaller European countries, however, is to use a set of indicators, rather than research evaluation, for the performance-based part of the institutional funding. This is the current situation in all Scandinavian countries. Bibliometric indicators of research performance are used in combination with other indicators representing external funding, doctoral dissertations and educational activity. Only the bibliometric indicators will be discussed in the following. They have been important drivers in the development of research information systems.

How performance-based funding can drive the development of research information systems is particularly evident from the experiences with the so-called "Norwegian Model" (Schneider, 2009; Ahlgren et al., 2012; Ossenblok et al.; Sivertsen & Larsen, 2012), which so far has been adopted at the national level by Denmark, Finland, and Norway, partly also by Belgium (Flanders) and Portugal, as well as at the local level by several Swedish universities. The model has three components:

A. A complete representation in a national database of structured, verifiable and validated bibliographical records of the peer-reviewed scholarly literature in all areas of research;
B. A publication indicator with a system of weights that makes field-specific publishing traditions comparable across fields in the measurement of "Publication points" at the level of institutions;
C. A performance-based funding model which reallocates a small proportion of the annual direct institutional funding according the institutions' shares in the total of Publication points.

Component B, the bibliometric indicator itself, is to our knowledge the first one to represent productivity across all fields of research in a balanced way.

In principle, the funding model in component C is not necessary to establish components A and B. The experience is, however, that the funding models in C support the need for completeness and validation of the bibliographic data in component A. Since the largest commercial data sources, such as Scopus or Web of Science, so far lack the completeness needed for the model to function properly, the bibliographic data are delivered by the institutions themselves through Current Research Information Systems (CRIS).

With regard to Component A, the four Scandinavian countries in focus here have chosen different solutions as a response to the need for data production through a CRIS system:

As *Denmark* introduced the Norwegian model as part of the funding model for its eight universities in 2009, most of them had already implemented the PURE system for local purposes. This is a commercial CRIS system that was developed in Denmark and is now delivered worldwide by Elsevier. All Danish universities use their own versions of PURE without integration at the national level. Instead, the bibliographic data are exported annually to a central database that has been designed specifically for the government to serve the need for component C.

Finland introduced the Norwegian model in 2015 after a process in which the universities chose not to replace their already installed local CRIS systems by the integrated national CRIS system suggested by the government. Instead, just as in Denmark, the bibliographic data are exported annually to a central database which has been designed specifically to serve the need for component C. At the local level, the Finnish universities continue to use *different* commercial or non-commercial local systems.

Norway had two different non-commercial CRIS systems in its higher education sector as the model was introduced in 2004. These systems were replaced by a fully integrated non-commercial national system in 2010 which is called CRISTIN (Current Research Information System in Norway). At the same time, the independent research institutes and the hospitals were invited to participate. Consequently, almost all research organizations in Norway's public sector now provide complete data for their scientific and scholarly publications. References to publications appear only once even if they have authors at more than one institution.

Sweden departed from the other Scandinavian countries in 2009 by deciding to use publications and citations from *Web of Science* as the only source of data for Component A (Sandström & Sandström, 2008; Flodström, 2011). Without the need for national data in component A, Sweden is so far an example of a lack of a driver for creating comprehensive current research information systems. The national system for data on scholarly publishing, which is called *Swepub*, has so far been incomplete and without standardized data. There are now detailed plans, however, for improving *Swepub* by integrating it into a new national CRIS system together with *Prisma*, the new application system of the Swedish Research Council. These plans are related to the description of a completely new performance-based funding model, FOKUS (Research Quality Evaluation in Sweden), which is inspired by the Research Excellence Framework in the United Kingdom (Swedish Research Council, 2015). The ambition is that the universities will not need to submit any information for the purpose of the evaluation only. The data collection for the evaluation will instead rely entirely on the national research information system, which will be running also for the universities' internal purposes. By late 2015, the new funding model in Sweden has not yet been politically sanctioned, but the process of improving *Swepub* has already started independently of the decision.

Since 2009, several Swedish universities have adopted the "Norwegian model" for internal purposes, thereby creating comprehensive and structured bibliographical data at the local level. The choice of the Norwegian model by individual institutions in Sweden is parallel to the decision in Flanders (Belgium) in 2008 to create a supplementary database, VABB-SHW, covering the humanities and social sciences with complete bibliographical data for the national funding model. This was a response to a political unrest in the humanities and social sciences because the funding model had been restricted to data from Web of Science since 2003 (Engels et al., 2012).

NordForsk, an organisation that facilitates and provides funding for research cooperation and research infrastructures among the Scandinavian countries, has recently initiated two related projects in which the countries will collaborate on further technical development of the national research information systems and on bibliometric analysis of the data. A mainstay of the information systems is a dynamic, global and standardized register of peer reviewed scholarly and scientific publication channels (journals, series, book publishers). The registers in each of the countries will be compared and developed into a Scandinavian standard register as part of the technical project, and there will be a division of labour connected to it. The standardization of registers is a prerequisite for cross-national comparison in the analytical project.

A parallel to the standardized registers of publication channels in Scandinavia is the European Reference Index for the Humanities and the Social Sciences (ERIH PLUS), which was created and developed by European researchers under the coordination of the Standing Committee for the Humanities (SCH) of the European Science Foundation (ESF). In 2014, responsibility for the maintenance and operation of ERIH was transferred to the Norwegian Social Science Data Services (NSD) where the Norwegian register of publication channels is hosted as well. The European register at NSD is called ERIH PLUS in order to indicate that it has been extended to include the social sciences. It has also become dynamic – suggestions for new journals can be made at any time. The criteria for inclusion have been made more objective, and the journals are no longer ranked. One of the aims of the new ERIH PLUS is to become a resource for the development of current research information systems with standardized and analysable bibliographical data.

Limitations and potentials of Current Research Information Systems from a bibliometric point of view

CRIS systems presently have two major limitations that may seem to make them completely uninteresting from a bibliometric point of view: They do not allow for international comparison or benchmarking, and they lack data on citations. However, they do allow for co-authorship analysis, studies of differences and trends in productivity, studies of research profiles and publication patterns, and they allow for text mining to the extent that the systems are connected to full text repositories. The values added by using CRIS systems for such analysis is on the one hand that they have *more complete* coverage of the scholarly literature than is found in the commercial databases, and on the other hand that the data are automatically *disambiguated* with regard to persons and affiliations and may thereby be *connected* to other data in the system.

In addition, citation analysis and international comparison is possible if the data are matched to data from Scopus and Web of Science. This happens almost automatically in the Norwegian CRISTIN system, because bibliographical records from the two external sources are imported into CRISTIN and validated there in order to facilitate the researchers' registration of the publications. This makes it possible to relate all Norwegian author names and institutional addresses directly to real persons and institutional affiliations in CRISTIN. This option not only takes away disambiguation problems, but also opens up for combining bibliometric data with other data representing researchers, activities and resources at

Norwegian research institutions. It also supports the Research Council of Norway with improved data for its research evaluation of disciplines on the national level. The data in CRISTIN are already used online by the Research Council instead of publication lists in the applications for funding.

Fruitful combinations of national CRIS data and international data from Web of Science in studies of scientific impact, productivity and mobility have already been illustrated in a few publications (Aksnes et al., 2011; Aksnes et al., 2013). I will give one further illustration here of how the relations between gender, age and scientific publishing can be studied by using connected CRIS data.

In the Norwegian CRISTIN system, gender, age and complete records of all peer-reviewed scientific publications is among the available information for each active researcher. We studied the productivity of 17,212 researchers (10,279 men and 6,933 women) aged 27-67 who published in 2011. Altogether, they contributed to 12,441 unique publications. There was no double counting if two or more researchers contributed to the same publication. Instead, publications with multi-authorship were fractionalized by the number of authors. *Figure 1* shows the result by presenting the women's share of among Norwegian researchers and their publication output in each one-year age cohort between 27 and 67.

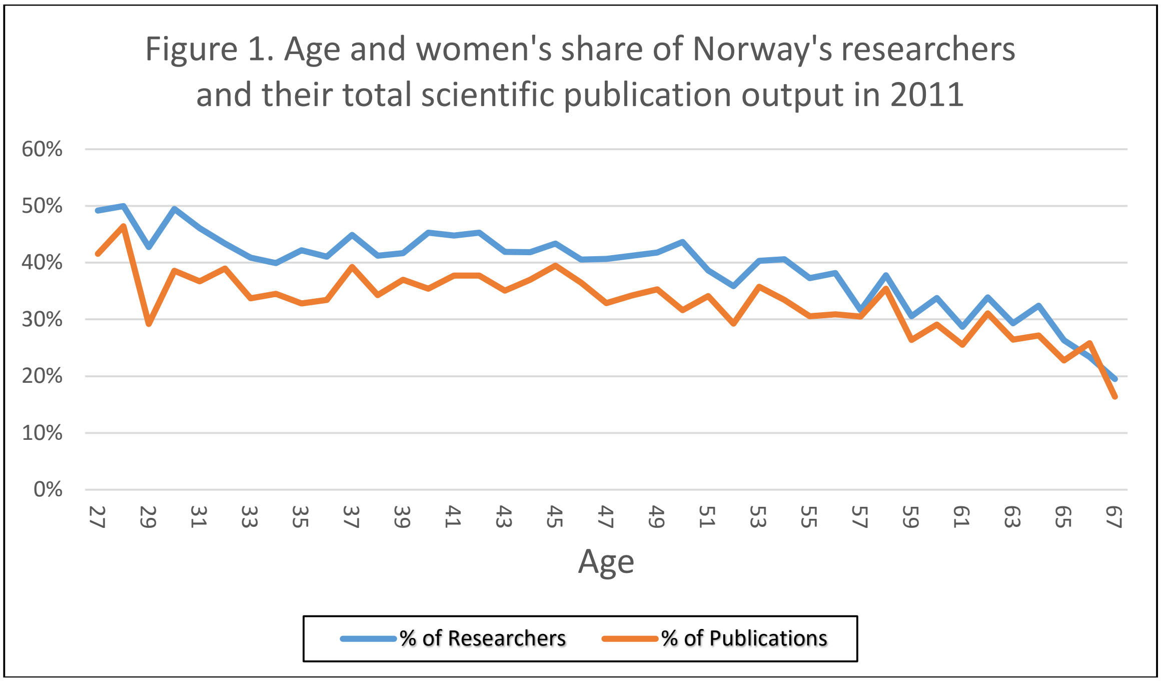

Figure 1. Age and women's share of Norway's researchers and their total scientific publication output in 2011. Based on data from CRISTIN, representing more than 17,000 active researchers working at 160 different research institutions in Norway.

We can see the gender gap decreasing as younger generations are recruited to research. We also observe that the difference in productivity between men and women is somewhat larger in the younger age cohorts. This is not a new finding. The same observation and its possible explanations have been studied more extensively in previous studies, e.g. by Kyvik & Teigen (1996) with the telling title "Child Care, Research Collaboration, and Gender Differences in

Scientific Productivity". That study, however, was based on a survey and interviews with relatively few researchers. *Figure 1* is based on complete data for all active researchers in a country. With the help of the CRIS system, we can see that the difference in productivity between men and women is so far consistent across all types of institutions (universities, university colleges, research institutes, hospitals) and across all fields of research (humanities, social sciences, health sciences and natural sciences). This could be an indication that gender equality in research is dependent also on external gender equality.

Conclusions

Well integrated and structured research information systems on the institutional or national level are being developed for several purposes, including local management and national funding. However, they seem to be promising also as data sources for studies of researchers and their activities, including bibliometric studies. The strength of these systems is connected to the completeness of bibliographical records, the automatic disambiguation of authors/persons and addresses/affiliations, and the possibility of thereby to connect with other data describing the researchers, their institutions and resources, and the outcomes of their research. It still remains to make this type of data comparable across countries, but recent developments in Scandinavia are pointing in this direction.

References

Ahlgren, P., Colliander, C., Persson, O. (2012). Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics*, 92(3), 767-780.

Aksnes, D.W., Rorstad, K., Piro, F., Sivertsen, G. (2011). Are Female Researchers Less Cited? A Large-Scale Study of Norwegian Scientists. *Journal of the Association for Information Science and Technology*, 62(4), 628-636.

Aksnes, D.W., Rorstad, K., Piro, F., Sivertsen, G. (2013). Are mobile researchers more productive and cited than non-mobile researchers? A large-scale study of Norwegian scientists. *Research Evaluation*, 22(4), 215-223.

Bahieu, B., Arnold, E., & Kolarz, P. (2014). *Measuring scientific performance for improved policy making.* Brussels: European Parliamentary Research Service.

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (2012). Changing publication patterns in the social sciences and humanities 2000-2009. *Scientometrics*, 93(2), 373-390.

Flodström, A. (2011). *Prestationsbaserad resurstilldelning för universitet och högskolor.* Stockholm: Ministry of Education and Research.

Hicks, D. & Wang, J. (2010). Towards a Bibliometric Database for the Social Sciences and Humanities – a European Scoping Project. Final Report on Project for the European Science Foundation.

Hicks, D. (2012). Performance-based university funding systems, *Research Policy*, 41(2): 251-261.

Kyvik, S. & Teigen, M. (1996). Child care, research collaboration, and gender differences in scientific productivity. *Science, Technology & Human Values*, 21(1), 54-71.

Ossenblok, T. L., Engels, T. C., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science – a comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). Research Evaluation, 21(4), 280-290.

Sandström. U. & Sandström, E. (2008). *Resources for Citations* (in Swedish: Resurser för citeringar). Stockholm: National Board for Higher Education.

Schneider, J. W. (2009). An Outline of the Bibliometric Indicator used for Performance-based Funding of Research Institutions in Norway, *European Political Science*, 8: 364–78.

Sivertsen, G. (2010). A Performance Indicator based on Complete Data for the Scientific Publication Output at Research Institutions, *ISSI Newsletter*, 6: 22–8.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, 91(2), 567- 575.

Swedish Research Council. (2015). *Research Quality Evaluation in Sweden – FOKUS.* Stockholm: Swedish Research Council.