S&T indicators as a tool for formative evaluation of research programs

Benedetto Lepori, *Centre for Organizational Research, Faculty of Economics, University of Lugano, via Lambertenghi 10a, 6904 Lugano, Switzerland*, blepori@unisi.ch

Emanuela Reale, *CNR-CERIS, Institute for research on firm and growth, via dei Taurini, 19 00185 Rome, Italy*, e.reale@cnr.ceris.it

1 Introduction

Ten years ago, Hariolf Grupp remarked that the literature dealing with the use of Science and Technology (S&T) indicators for program evaluation was limited (Grupp 2000) and mostly focused on bibliometric indicators. An assessment of the current situation would come to similar results. Despite the longstanding debate on the increased use of quantitative indicators to complement qualitative approaches (Roessner 2000; European Court of Auditors 2007), and despite the rapid development of the S&T indicators domain, as well as of their use in other domains of evaluation like funding agencies and research institutions (Campbell 2003;Whitley and Glaser 2007; HEFCE (Higher Education Funding Council for England) 2007), there are few examples of the systematic use of indicators in program evaluation. Reviews by expert panels and impact assessment using large-scale surveys on projects and grant recipients have been until now the most widespread approaches (Georghiou and Larédo 2006), while in the most well-known evaluation manuals (Ruegg and Feller 2003; IPTS 2002) S&T indicators make a rather minor appearance.

A recent study on European Framework Programs (EU FPs) also comes to deceiving results (Proneos, OST, NIFU-Step and CERIS-CNR 2009): the existing indicators are hardly suitable to provide evidence for most of the envisaged evaluation questions, while the designing of *ad hoc* indicators would be a resource-consuming process, in many cases severely limited by the availability of data.

One could argue that S&T indicators are not well suited to this purpose and that the debate should be discarded as an outcome of the fascination with numbers (Roessner 2000), accepting the inherently fuzzy and subjective nature of program evaluation (Feller 2007). This argument is supported by two methodological problems regarding the use of S&T indicators for evaluation, namely *demarcation* – it is difficult to attribute research results to a specific funding program – and *time-lags* – most outputs become visible after the end of the program (Grupp 2000).

However, our thesis is that these limitations are also related to underlying conceptions of S&T indicators as exact measures, providing quantitative answers to questions, as well as to a largely summative conception of program evaluation, where the main goal is to objectively assess the degree of achievement of program goals. Instead, we argue that a broader and richer contribution of indicators can be envisaged within a conception of program evaluation oriented towards formative evaluation – learning lessons from the past in order to improve future program generation (Thoenig 2000) –, as well as within a socially-constructed discourse among the stakeholders involved in program operations, rather than as a top-down process ordered by the program owner.

We further argue that, while this conception is of broader significance for evaluation activities as a whole, since it is related to constructivist (Guba and Lincoln 1989) and theory-based approaches to evaluation (Stame 2004), it is particularly significant in the case of research programs, as they are typically multi-actor

and highly distributed settings, with multiple and in some cases conflicting goals. While the recognition of these characteristics has driven research program evaluation towards formative and constructivist approaches (e.g. see Kuhlmann 1998, Molas-Gallart and Davies 2005), the implications for the use of S&T indicators have been hardly developed.

In the following, we develop our argument in three steps. In section 2, we provide introductory notions on the epistemological status and development of S&T indicators, while also highlighting their strength and weaknesses for the purposes of program evaluation. In section 3, we develop our core argument on research programs as multi-actor interaction spaces and on evaluation as a mediation process among the actors' interests and needs. In section 4, we derive a framework for the use of S&T indicators in evaluation, we highlight their potential, and we draw practical implications for how to handle their integration in evaluation processes. We conclude with a short discussion of more general implications.

2 The epistemological and social foundations of S&T indicators

The field of science and technology indicators was first developed through pioneering work during the '60 at the OECD – focusing on measures of research input and especially R&D expenditures and human resources (Godin 2005) – and by scholars interested in systematic ways of measuring scientific production (Martin and Irvine 1983). During the last three decades, the field has broadened and professionalized through the definition of methodological standards, the development of large-scale databases, as well as the creation of specialized centers (Lepori, Barré and Filliatreau 2008).

While most of the works on indicators have adopted a positivistic view, assuming them to be objective measures of properties of S&T systems, a small body of literature, rooted in insights from sociology of measurement and statistics (Desrosières 2008; Porter 1995), has developed a conceptualization of them as socio-cognitive constructs, which reflect underlying assumptions on the world and thus depend on the values and interests of the actors developing them (Godin 2005). As shown in the case of scientific performance of countries, even the simplest indicators provide different results depending on the assumptions on how they should be constructed and normalized (Barré 2001).

Does this mean that indicators are nothing more than instruments to legitimize decisions already taken on different grounds and to impose the power of the State on society, as some sociological works suggest (Godin 2005)?

We consider this position too extreme. Very much like all kinds of measures, indicators depend on the underlying representations of reality and on the actors' frames of reference; nevertheless, they are not arbitrary. On the contrary, a socio-constructivist approach highlights some central criteria to assess their quality and reliability, including the soundness of the theoretical framework and its applicability to the specific situation, the extent to which indicators have been validated through other kinds of measures and sources of errors have been detected and, finally, the transparency concerning the objectives and value choices of the actors involved.

Two central implications are that indicators are *context-specific* and debatable. They need to be interpreted taking into account their specific context of usage, limitations of applicability, and errors of measurement. Moreover, the actors might provide diverging interpretations of the same indicator, based on their frames of reference and goals. In other words, indicators do not provide answers to policy and evaluative questions; on the contrary, they are tools to structure and foster debate in multi-actor social spaces (Barré 2004). As we shall see, this makes them a potentially relevant tool in the framework of participative and moderating conceptions of S&T evaluation (Kuhlmann 1998).

Since there is considerable fuzziness over where to draw the distinction between data and indicators, it is useful to distinguish among: *descriptors* – describing some aspects of reality without leading to further interpretation –, *markers* – used in place of quantities that cannot be measured directly, for example patents as a marker of technological outputs –, and *indicators* – explicitly building connections between quantities and non-observable properties. While descriptors are largely adopted in program evaluation – e.g. basic statistics on selection, participation, and results –, our focus is especially on the use of more theory-based indicators, which allow us to produce statements on deeper issues like quality of research, societal impacts, or fairness of selection.

2.1 A short overview of the field of S&T indicators

A consultation of the most recent handbook of quantitative S&T studies (Moed, Glänzel and Schmoch 2004) reveals an impressive range of available indicators, as well as a broadening scope of their usage. The recent Innovation Union Competitiveness report published by the European Union also documents the number of S&T indicators available at the European level (European Commission 2011).

A simple categorization rests on the so-called linear model of innovation, which has driven the development of S&T indicators from the '60 onwards (Godin 2005), thus leading to a distinction among input, output, and impact indicators (which is still largely adopted also in S&T evaluation; Molas-Gallart and Davies 2005). Within the domain of input indicators, the OECD has taken on a leading role by developing standards for measuring research expenditure and funding (Frascati Manual; OECD 2002), innovation (Oslo Manual; OECD 2005), as well as human resources devoted to science and technology (Canberra Manual; OECD 1995).

Indicators on scientific outputs have focused on the analysis of scientific publications through bibliometrics, probably the most developed and fastest growing domain in S&T indicators overall (van Raan 2004), which is increasingly characterized by a plurality of methods and data sources (e.g. to better cover social sciences and humanities; Nederhof 2006). Concerning technological outputs, patens have become a standard indicator, thanks to the availability of large international databases (Breschi and Lissoni 2004). The broader field of transfer of research activities to economy and society ("third mission") remains however rather poorly covered (Gulbrandsen and Slipersaeter 2007). In the more general area of economic impact, considerable progress has been made regarding innovation indicators, through data collection from the Community Innovation Survey (Bloch 2007). A few additional indicators on economic impacts are routinely used in analyses like the European Innovation Scoreboard; these include: SMEs introducing product or process innovation, employment in high-tech firms and services, manufacturing and service exports, net-to market and net-to-service sales. On the contrary, indicators measuring the social and environmental impact of research activities are probably the least developed domains(Luukkonen 1998).

A relevant tendency for evaluation purposes is the shift from national aggregates – comparing national innovation systems and their performance – to indicators at the level of individual actors – e.g. funding agencies (Lepori, Dinges, Reale, Slipersaeter, Theves and Van den Besselaar 2007) or higher education institutions (Bonaccorsi, Daraio, Lepori and Slipersaeter 2007) – , as well as to indicators characterizing the position and linkages of individual actors – the so-called *positioning indicators* (Lepori, Barré and Filliatreau 2008). Interesting applications for research evaluation are studies mapping the structure of scientific fields combining different types of information, like bibliometrics, funding projects, and policy documents (Merkx and Van den Besselaar 2008). From a technical point of view, the development of software tools for the indexing of large corpora of heterogeneous documents allows this approach to be extended beyond the search strategies on databases (Grupp 2000), including also textual documents like project descriptions and policy documents and avoiding taxonomical problems (Mogoutov and Kahane 2007).

2.2 Lessons from past experiences

To which extent have concepts and methods developed in the broader field of S&T indicators been applied in the practice of program evaluation? Has the situation changed significantly in the last decade since Grupp's review (Grupp 2000)?

At least for what concerns the European context, a few examples show that the situation has not changed significantly in the last ten years – a fact that is confirmed by the low visibility of Grupp's paper, which has received very few citations both in the Web of Science (1) and in Google Scholar (6).

A meta-analysis of the evaluations related to EU FPs (at both the European and national level) shows that quantitative approaches and S&T indicators have hardly ever been used and most evaluations rely on other methods, like experts, participants' questionnaires, and reviewing of official documents (Arnold, Clark and Muscio 2005). Both the ex-post evaluation of FP6 and the interim evaluation of FP7 are essentially based on panel review, supported by extensive statistics and data on participation patterns, as a well as by a number of impact assessment studies; S&T indicators are virtually absent in both reports. When looking at the larger set of impact studies produced in this context, most of them are based on participants' surveys or interviews.

However, two relevant exceptions stand out. First, a study on the bibliometric profiling of FP participants provides an analysis of the overall publication output and impact of lead scientists participating in FPs (identified through a survey of FP project coordinators). This study shows that lead scientists in FPs have a publication and citation performance, which is higher than that of their counterparts within their scientific community (Technopolis and OST 2009). Second, a study using FP participation data provides relevant insights into the formation of the European research network, as well as into the emergence of a backbone of participants in FPs (AVEDAS, CWTS, FAS.Research and Stockholm School of Economics 2009). These two studies display some relevant trends in the use of S&T indicators for evaluation, namely a shift from projects to participants, the combination of participation data and bibliometric indicators, and a stronger focus on linkages and structuring effects.

A further attempt at using S&T indicators more extensively, especially to map the European research landscape and to measure the outputs of previous FPs, was made in the ex-ante impact assessment of FP7 (Delanghe and Muldur 2007). The authors explain that the improvement of ex-ante impact assessment requires more rigorous and comprehensive collection of data on applicants and participants, as well as on outputs and impacts, which can be achieved through new reporting techniques and improved methodologies for the use of bibliometric or innovation data (i.e. from the Community Innovation Survey). Moreover, a new effort in the field of S&T indicators is suggested, with a shift from inputs to flows (collaboration patterns, co-publications, co-patenting, etc.).

Concerning the measurement of project results, bibliometric indicators are reasonably well-established, especially for programs in domains that are thoroughly covered by international databases (Technopolis and OST Paris 2004, van Raan 2004), while technological indicators are much less developed. If we carry out a mapping exercise concerning the evaluation practices and methods adopted by several Funding Agencies in Europe to assess strategic issues and impact, research fields and disciplines, and funding programs, we can clearly see that output indicators – such as papers, PhDs, and patents – are always considered in the analysis, although their role is not yet a prominent one, nor are they fully integrated in the evaluation design. Moreover, the analysis shows an ongoing shift in the main rationale for evaluation towards formative evaluations (Reale, Inzelt, Lepori and van den Besselaar 2011).

As to the impact, an attempt at using quantitative evidence, comparing existing studies and practices of ex-post assessment of impacts from basic research, suggests that indicators should distinguish between the scientific community and the users outside the scientific community in order to capture the effective impact of programs, while the outcome of research should be used as a proxy measure of societal impact (Kanninen and Lemola 2006). The basic choice concerns how to collect the data used for the indicators: if the data are gathered from the organisations, the outcome cannot be attributed to a particular research program; if the data are gathered from the project level, the impact produced after data collection remains unknown, and the non-linear effects of research cannot be understood. One should select the best approach on the basis of the evaluation aims: the former approach aims at assessing the impact regardless of which project has generated it; the latter focuses specifically on the effects of one instrument, regardless of how incomplete the general picture might be.

We believe that there are some valid reasons to overcome the traditional scepticism displayed by evaluation specialists towards the use of quantitative indicators (ad least concerning research programs; Cozzens 1997). This would lead to an enhanced contribution to program evaluation.

First, as explained above, the range of available S&T indicators has greatly increased in the last few years, and many of them could be profitably applied to program evaluation too. Network and linkages indicators are a case in point; in fact, the recent Innovation Union Competitiveness Report (European Commission 2011) makes extensive use of said indicators to discuss the structural impact of the FP on the European Research Area (see also Heller-Schuh, Barber, Henriques, et al 2011, Scherngell and Barber 2011). Second, the ongoing debate within the evaluation community on how to trace the outcomes and impacts of research programs (Rogers and Jordan 2011) shows that the experts' judgement and the participants' self-assessment alone will not suffice, but some kind of quantitative measurement will be required. The inability to address this issue might result in delegitimizing the evaluation results and in reducing their political impact.

However, we argue that, since indicators are not purely objective and technical instruments, their role cannot be properly investigated without dealing with deeper issues concerning the nature of program evaluation itself. This will lead to a more realistic representation of how indicators can contribute to evaluation, while also broadening their focus from the measurement of results and impacts towards the assessment of program assumptions and operations – two questions that are largely neglected in summative evaluations.

3 Conceptualizing research programs and their evaluation

Traditionally, evaluation was dominated by a summative approach, whose main objective was to measure the extent to which program goals were achieved. Accordingly, it was assumed that program owners designed their methods rationally, based on some underlying intervention logic, which was not to be questioned by the evaluation itself (Stame 2004). Hence, evaluation approaches focused on methodologies to evaluate outputs and impacts against a non-intervention benchmark, like randomized experimental designs. The quest for quantitative indicators largely follows this paradigm.

New evaluation approaches have been developed in the last decades in order to cope with the complex reality of most programs and with social demand for participation. These include responsive evaluation (Stake and Abma 2005, Abma 2006), constructivist evaluation (Guba and Lincoln 1989), as well as theory-based evaluation (Weiss 2004; Stame 2004). While it is not the purpose of this paper to discuss these approaches in general, we wish to highlight some issues that apply specifically to research funding programs.

3.1 Programs as socially-constructed actor spaces

Research programs are one of the main tools adopted in research policy in order to achieve broader goals such as scientific excellence and social relevance; they provide financial incentives to research groups to direct their research agenda towards public goals; competition for resources is created through calls for proposals – hence, the groups striving to receive funds should align their research priorities to those of the principal (Braun 2003). Nevertheless, research programs are an extremely differentiated domain for what concerns their objectives, the rules for the selection of proposals, the type of output expected, and the contractual relationship between funding bodies and research groups. Responsive mode programs managed by research councils provide funding for basic research on topics proposed by the researchers, allowing for great freedom in the research performed, whereas contract research explicitly requires the delivery of outputs defined at the contractual level. These differences also have an impact on the extent to which a research program should be considered a tool to implement public policies or a socially-constructed interaction space between policy and science (and, accordingly, on the balance between summative and responsive evaluation).

A first relevant characteristic of research programs is their *high level of delegation*, complemented by *a distinctive principal-agent structure* (Braun and Guston 2003), whereby the funding agency allocates money on the basis of the proposals it receives, but the decisions about how to carry out the projects are largely left to the research groups. Delegation is a well-known phenomenon in public programs – related to the diffusion of new public management practices – and a well-known issue in evaluation (Stame 2004). It has distinctive features, both because research is an inherently risky undertaking whose success depends on the creativity of the performer and because controlling the results and measuring their quality is difficult. For what concerns research programs, delegation is not purely good management practice, but it is also related to an understanding of research systems as distributed intelligence settings, where there cannot be a single actor (e.g. the State) defining the goals and steering the system centrally (Kuhlmann, Boekholt, Georghiou, et al 1999).

Second, today's research funding systems are characterized by the *coexistence of different funding schemes*, which tend to overlap in terms of the research they support, e.g. thematically-oriented vs. investigator-driven programs or programs managed at different institutional levels (national vs. European). While in other domains this is considered a non-optimal situation and integrated programs have been promoted, for what concerns the research domain differentiation of funding schemes is seen as a sound policy to address partially conflicting goals and to encourage competition. Current research funding systems can be better understood as a market space allowing coordination between the *funders* – providing resources related to specific policy goals – and the *performers* – providing abilities and competences –, rather than as a hierarchically organized implementation system (Lepori 2011). In STI policy evaluation, this has led to the emergence of approaches like systemic evaluation (Arnold 2004) or intelligent benchmarking (Guy and Nauwelaars 2003).

Third, almost *all research programs are, to a certain extent, jointly designed and implemented together with their beneficiaries*. Co-designing is relevant since programs owners do not usually have enough competences to identify the most promising research areas. Joint management is needed since the competences for evaluating and selecting best-quality research are available from the researchers themselves, hence their involvement in evaluation and selection committees. Accordingly, the designing and implementation of research programs can be best understood as a negotiation process between the State – which holds the resources and strives to achieve policy goals – and the research community – which possesses information and competences and whose members strive to pursue their own research agenda.

Obviously, as the State provides almost all the resources, a certain level of accountability to the original policy goals is required (implying that summative evaluation will always be relevant); however, what matters in this context is a conception of research programs as complex interaction spaces ("boundary objects"; Guston 1999, Klerkx and Leeuwis 2008) among largely autonomous and strategic actors belonging to different social spheres ("policy", "society", "science"). Their function is to negotiate program designing and operations rather than to implement ex-ante defined policy objectives. As there are multiple spaces in which the interaction among the same actors – the State, funding agencies, performers – takes place concurrently, programs are inherently open settings and their working cannot be adequately understood without taking into account the interaction with other funding schemes.

3.2 From program characteristics to evaluation approaches

Some scholars have developed the implications of this perspective for research policy and programs evaluation (Kuhlmann 2003); yet, these are slow to penetrate into evaluation practices, which still tend to adhere to a more linear understanding of how research works (Molas-Gallart 2005).

First, if programs are implemented by largely autonomous actors, evaluation has to start from their needs and from how they interpret the program goals, since these features are likely to be more relevant for program operation than the *ex-ante* goals stated in official documents; like in responsive evaluation, evaluation objectives and questions need to be constructed together with the program participants, rather than decided by the program owner alone. At the same time, how these actors respond to program measures and (financial) incentives determines the program outcomes and impacts, a lesson endorsed by studies focusing on behavioral additionality (OECD 2006). This means that the evaluation has to consider which action theory underpins the actors' behavior and to which extent the programs influence it - not solely by measuring outputs and impacts, but also by explaining why and how these were generated.

Second, if programs have inherently open and fuzzy boundaries, problems of demarcation are fundamental and cannot be addressed only through sophistication of analytic methods. Trying to attribute outcomes and impacts to a specific program is, to some extent, a contestable exercise and provides less and less meaningful results when moving towards the broader impact on science and society. Two promising approaches are focusing on participants rather than on projects and investigating interactions as sources of broader impact (Molas-Gallart 2011, Spaapen and van Drooge 2011). Both approaches are examples of the shift towards theory-led evaluation. Assuming that research groups are able to develop long-term strategies, if a research program is able to shift the participants' agenda, there is a chance that long-term results will be generated. Moreover, assuming that creating linkages and interactions among the stakeholders has a social impact, if a research program manages to strengthen these ties, this is likely to generate long-term impacts, which will hardly be measurable immediately after the end of the program itself.

Third, if programs are co-designed by a complex web of actors, addressing the questions and needs of the program owner alone is not sufficient. Rather, evaluation should become a source of strategic intelligence, enabling negotiation among the actors, making them more aware of the implications of different program designs, as well as of the interests and strategies of the involved actors. In this perspective, evaluation should not produce pre-defined recommendations and conclusions but rather an agenda for the negotiation among the actors, highlighting critical issues, open questions, and possible choices to be made (Kuhlmann 1998). As foreseen by socio-constructivist evaluation (Guba and Lincoln 1989), constant communication with the program actors throughout the whole process becomes at least as important as the final evaluation results.

4 A framework for the use of indicators in program evaluation

Our discussion of program evaluation draws on concepts introduced in section 2, like the value-laden nature of indicators and their status as instruments for the social debate. This justifies our claim that a broader conception of evaluation provides room for a more extensive usage of S&T indicators, alongside their traditional function of measuring outputs and impacts in summative evaluation.

4.1 Indicators as contributions to the evaluation debate

The formative and constructivist conceptions of evaluation imply a shift in the timing of the process, from an approach in which evaluation is performed at the end of the program towards a continuous process throughout the whole program life, in which evidence from evaluation activities – indicators, surveys, experts' assessment – is debated and evaluated by program actors when it becomes available. The current evaluation setting of the European Framework programs including ex-ante, mid-term and final evaluation is an example of this approach (Molas-Gallart and Davies 2005).

This process approach potentially facilitates the use of S&T indicators, as it provides room for interactive designing and refinement during the program life. When the main evaluation questions are constructed at the beginning of a program, it is crucial to determine which indicators are needed and how these can be designed, while also identifying the data that should be collected during the program's life. Then, at a later stage, a first set of indicators should be provided and debated with the program actors, to test in advance their feasibility, reliability and relevance, and to devise possible refinements. This interactive approach is ideal not only for technical reasons, but also because the nature of S&T indicators requires them to be co-constructed together with the involved actors, in order to have them accepted and to allow for effective impact on the decision-making processes (Barré 2004).

Conceiving evaluation as a process also implies a different use of S&T indicators, focusing on their ability to promote communication and the learning processes rather than on objectivity and precision.

As argued in section 2, this is their main strength. By adopting a strong stance on the characteristics of reality and on value choices, they are able to simplify an overly complex reality to few numbers, which can be produced with a reasonable amount of effort. As such, they can raise questions, highlight assumptions of the participating actors, and provide counterevidence to established positions. For instance, trying to use bibliometrics to measure the output of EU FPs was a resource-consuming exercise (Technopolis and OST Paris 2004), while using the simple impact factor is more feasible and likely to stir some relevant debate on questions such as: are the best scientists in Europe participating in FPs? Are some of the best publications in a field related to EU-FP funded research?

This information is not meant to provide causal demonstration, but rather to act as input for broader evaluation debate, so that it can be compared to other evidence and evaluated according to different perspectives, as well as to its reference context – for example what would normally be expected at that stage of development of a program. In this respect, timeliness becomes a central criterion. Many ambiguities in the use of indicators can be ascribed to misplaced expectations on their precision – which lead to unrealistic recommendations on collecting additional data rather than on pragmatic strategies to make the best of available information – and to attempts at using numbers to draw conclusions rather than to promote the evaluation debate.

Of course, the value of the adopted indicators will also depend on their *reliability*, i.e. confidence that they can actually measure what they should measure, as well on their *relevance* in relation to the evaluation questions. The former is related to the soundness of the theoretical assumptions, the quality of the data sources, but also the robustness and transparency of the production processes (Barré 2004). Hence, it is

fundamental that indicators are provided by specialists, who take stock of the state of the art in the field and are aware of their limitations. Relevance is related to a deep understanding of the questions being raised and, thus, to good integration in the evaluation process.

4.2 Overcoming delimitation problems: from project to participants

Our discussion of research programs supports the idea that program participants, rather than funded projects, are the central analytical unit to be observed, as they are able to develop long-term strategies and research agendas and to combine available resources in order to pursue them (Latour and Woolgar 1979, Joly and Mangematin 1996). Analyzing the strategies of research groups and contrasting them with those of other groups not participating in the program is relevant for different purposes: *ex-ante*, in order to foresee the results which might be expected and to identify critical limitations, e.g. the lack of interaction with societal stakeholders; during the program, in order to understand the groups' action logic and take it into account during implementation – e.g. to favor new entrants if there are few groups in a domain displaying collusive behavior; finally, in evaluating program results, since changes in participants' profiles, strategies and linkages can be considered a proxy of the program's long-term impact.

Research groups are also the natural unit of analysis in which S&T indicators are produced, as problems of delimitation are by far less serious than for projects and these indicators are routinely produced for other purposes, like institutional evaluation. They include indicators on scientific production - both publication counts and impact factors (van Raan 2004) - , as well as on technological production, based especially on patents (Grupp, Schmoch and Kuntze 1991). Limitations of the former relate to the poor coverage of social sciences and humanities (Nederhof 2006), but they can be overcome, to some extent, by combining heterogeneous data sources, while limitations concerning patents relate to the lack of a clear measure of their value, but also to identification problems in the public sector related to regulations on intellectual property. New methodologies based on the inventor's name are addressing this issue (Lissoni, Llerena, McKelvey and Bulat 2008).

A second category of indicators makes it possible to map the activity profiles of research groups (Larédo and Mustar 2000) and their trajectories over time (Braam and Van den Besselaar 2010). These features are highly relevant as they allow considering, *at the same time*, changes in the activity dimension of groups as well as their synergies and interactions, breaking with a tradition measuring separately the different types of outputs and impacts of programs.

A third category relates to linkages and cooperation both among research groups and with other relevant stakeholders. This category includes measures of scientific cooperation using co-publications data (Glänzel and Schubert 2005), science-technology linkages through patent citations analysis (Tijssen 2001), but also broader indicators of links with society, such as media presence, interlinking of websites, visibility in data sources, such as Google Scholar, which are not exclusively addressed to an academic audience. There are very valid reasons to focus on interaction indicators: first, awareness that scientific and technological productivity is to a large extent related to the level of cooperation and networking; second, evidence that 'productive interactions' between research groups and social stakeholders are a central mechanism to promote the social impact of research. Hence, measuring these characteristics might anticipate long-term impacts that are not yet measurable (Molas-Gallart and Davies 2005).

Gathering information on participants requires the availability of a database, allowing for unambiguous identification of participants at an aggregate level, which is relevant for evaluation across the whole program life (and possibly beyond). A database is necessary also for a fine-grained analysis of participation patterns, to administer participants' surveys, and to match participants with indicators from other sources (e.g., bibliometric data). Unfortunately, in many cases participants' databases are designed for

management purposes and are hardly usable for evaluation purposes. In FP7, for instance, the participants' databases list legal entities (although they also contain information on participating research units), which represent a problematic choice for evaluation purposes, since participation takes place at the unit level and legal entities have different meanings in each country. The consequence is that most evaluations require spending a large amount of time cleaning the available data (Delanghe and Muldur 2007) and compiling list of participants in order to administer surveys.

4.3 A broader set of evaluation questions

While indicators have traditionally been used mostly for the purposes of output and impact measurement, they can prove more useful in shedding light on two important matters, namely checking program assumptions and operations (Grupp 2000).

As for the assumptions, the formative perspective implies an open conception of research programs, embedded in a broader environment and largely meant to transform their features rather than to produce direct results. This implies that the correctness of the assumptions will be as important for the success of the program as its operations, and thus this question takes on a central role in evaluation.

Indicators can be supportive in two ways. Aggregate indicators at the national or sectorial level – research expenditure, impact factors by field, indicators on patents – provide a broad overview of how a research and technological sector is evolving and analyze to which extent gaps identified when launching a program have been addressed. Indicators at the group level can be used to produce maps of the actors in a field and of their networks of relationships, identifying core actors and critical linkages and providing rationales for the weakness of a research field, including excessive fragmentation, but also oligopolistic power and lack of competition; thus, they can lead to different types of program designing.

For what concerns the characterization of program operations, its relevance lies in the fact that it is not only a technical but also a strategic issue. A program might fail its objectives if it is not able to involve the right participants and to build the envisaged linkages, but also if the stakeholders and research actors are not integrated in the designing phase. The evaluation process has the aim to provide evidence on issues like the quality of communication, the functioning of the proposals' selection process, and the level of subscription to program goals. Hence, a shift from evaluation of administrative processes towards indicators characterizing the *participation of actors in program designing and evaluation processes* is suggested. Besides standard indicators on efficiency – like time for decisions and effort required to write a proposal –, useful indicators can be built from program reporting, including participation in committees and demography of evaluations, participation of actors in the selection process, success rates, and demography of proposers and participants. A *common system for the logging of administrative events* to analyze program management and administration will be required for this purpose.

4.4 Operational implications

The previous discussion has highlighted some organizational requirements needed to fully exploit the potential of S&T indicators for the purposes of program evaluation – most of which comply with good practices in organizing evaluations overall.

The first aspect is that advanced planning is extremely important, as collecting data, designing indicators, testing, and validation are processes that require time – especially in order to exploit the potential of indicators customized for the specific context of the evaluated program. Hence, the potential use of S&T indicators is an issue that should be discussed in the early phases of the program's life, when the overall evaluation goal, its timeframe and available resources are also defined. These requirements are similar to

those of other evaluation methods, like surveys, but in some cases evaluation specialists seem to adhere to a conception of indicators as off-the-shelf tools.

Second, indicators are highly dependent on the availability of data sources. While for some of them, like those used in mapping exercises, existing data can be used, other indicators rely on the availability of data on program activities, participants, and project results; thus, a well-organized system of program data collection becomes essential (Ruegg and Feller 2003). Key priorities in this respect are a well-structured participants' database, a system for the logging of administrative events and, finally, a system of *participants' and projects' surveys* to track the scientific and technological output and to collect the participants' opinions and feedback on management and impacts.

Third, indicators production is a highly specialized activity requiring specific competences, in conceptual terms, methodology and technical treatment of data, as well as resources in terms of manpower and infrastructures. We do not suggest in this context that all evaluation exercises should include a wide use of S&T indicators and provide significant resources for this purpose; rather, the choice about when to use them should be dictated by pragmatic considerations concerning their added value vs. issues of availability and effort, and there should be some reasonable relationships between goals and resources invested.

Finally, the use of indicators in evaluation heavily relies on the evaluators themselves having a clear idea of what S&T indicators are, of their potential contribution, and of the requirements for their designing and production. Accordingly, we argue that basic competences concerning S&T indicators should be an integral part of the training of evaluators specialized in research and technological programs, just like other methods such as panel reviews or surveys. Providing technical information on which indicators are available is, of course, very important, but even more important are the foundations concerning their epistemological status, the principle of their social construction together with the users, and the distinction between designing and production. All these topics can be easily related to recent approaches to evaluation overall – i.e. theory-based, responsive, and constructivist evaluation.

5 Discussion and conclusions

In this paper, we have shown that S&T indicators are powerful tools for understanding and interpreting research and development. While their use has traditionally been limited by serious constraints (i.e. demarcation and time lag), which have impeded the emergence of systematic work on their designing and production, we have argued that there are good reasons for a broader use of S&T indicators in future evaluation activities.

This is justified by parallel developments in both fields. The field of S&T indicators has become significantly differentiated in recent years, shifting from the production of standardized (national) aggregates to *positioning indicators*, characterizing individual actors and their linkages within the innovation system. This conceptual development is supported by the increased availability of data sources, new data retrieval techniques, and advancements in the tools for data exploitation and analysis, which make it possible to move towards customized and contextualized indicators. At the same time, the diffusion of theory-led evaluation approaches, focusing more on participants and interactions than on projects, makes it possible to envisage a much broader role for S&T indicators, which fits their status better than the requirement of providing objective answers to evaluation questions. We thus highlight the advantages of using S&T indicators when a formative perspective of evaluation is adopted and when programs are conceived as social constructions and actors' spaces rather than as tools for implementing public policy objectives.

Two final remarks should be added. On the one hand, indicators can contribute to the general debate about the validity and the advantages of public investment in R&D. In this respect, developing indicators

under a formative learning-oriented program evaluation approach must be seen as a complement to the summative efforts that contribute to feeding the "agenda for negotiation" (Kuhlman, 2003) among the different actors involved (policy actors, intermediaries, and performers). On the other hand, following a socio-constructivist approach, indicators can be easily integrated into comprehensive designs for the broader field of program evaluation, improving the ability of external evaluators to understand the value and potential of the complex research initiatives they have to assess (Trochim, Marcus, Masse, Moser and Weld 2008).

Finally, we wish to highlight the importance of preserving the differences in conceptual assumptions on what policy implementation is and what evaluation is, as well as the fact that good evaluation must be carefully related to the specificities of its context of application. Both these assumptions are relevant when the designing and validation of R&D indicators are concerned.

6 Acknowledgments

The authors acknowledge support from the European Commission under the contract 30-CE-1232670/00-40, as well as contributions from their colleagues involved in this contract (Michael Braun, Philippe Larédo, Stig Slipersater, Aris Kaloudis, Ghislaine Filliatreau, OST). Preliminary versions were presented at the European Forum on Research and Development Impact Assessment (EUFORDIA), Prague, February 2009, at the Triple Helix Conference, Glasgow, June 2009, as well as at the ENID/STI Indicators Conference, Paris, March 2010. Finally, the authors would like to acknowledge useful advice from three anonymous referees.

7 References

Abma, T., (2006). The Practice and Politics of Responsive Evaluation. *American Journal of Evaluation, 27 (1)*, 31-43.

Arnold, E. (2004). Evaluating research and innovation policy: a systems world needs systems evaluation. *Research Evaluation, 13(1)*, 3-17.

Arnold, E., Clark & Muscio (2005). What the evaluation record tells us about European Union Framework Programme performance. *Science and Public Policy, 32(5)*, 385-397.

AVEDAS, CWTS, FAS.Research & Stockholm School of Economics (2009). *Structuring Effects of Community Research – The Impact of the Framework Programme on Research and Technological Development (RTD) on Network Formation* Brussels.

Barré, R. (2004). S&T indicators for policy making in a changing science-society relationship. In H. F. Moed, W. Glänzel & U. Schmoch(Eds.) *Handbook of Quantitative Science and Technology Research.* (pp. 115-132). Dordrecht: Kluwer Academic Publishers.

Barré, R. (2001). Sense and nonsense of S&T productivity indicators. *Science and Public Policy, 28(4)*, 259-266.

Bloch, C. (2007). Assessing recent developments in innovation measurement: the third edition of the Oslo manual. *Science and Public Policy, 34(1)*, 23-34.

Bonaccorsi, A., Daraio, C., Lepori, B. & Slipersaeter, S. (2007). Indicators on individual higher education institutions: addressing data problems and comparability issues. *Research Evaluation, 16(2)*, 66-78.

Braam, R. & Van den Besselaar, P. (2010). Lyfe cycles of research groups: the case of CWTS. *Research Evaluation, 19(3)*, 173-184.

Braun, D. (2003). Lasting tensions in research policy-making - a delegation problem. *Science and Public Policy, 30(5)*, 309-321.

Braun, D. & Guston (2003). Principal-agent theory and research policy: an introduction. *Science and Public Policy, 30(5)*, 302-308.

Breschi, S. & Lissoni, F. (2004). Knowledge networks from patent data. In H. F. Moed, W. Glänzel & U. Schmoch(Eds.) *Handbook of Quantitative Science and Technology Research* (pp. 613-644). Dordrecht: Kluwer.

Campbell, D. F. J. (2003). The evaluation of university research in the united kingdom and the netherlands, germany and austria. In P. Shapira & S. Kuhlmann(Eds.) *Learning from Science and Technology Policy Evaluation* (pp. 98-131). Cheltenham: Edward Elgar.

Cozzens, S. E. (1997). The Knowledge Pool: Measurement Challenges in Evaluating Fundamental Research Programs. *Evaluation and Program Planning, 20(1)*, 77-89.

Delanghe, H. & Muldur (2007). Ex-ante impact assessment of research programmes: the experience of the European Union's 7th Framework Programme. *Science and Public Policy, 34(3)*, 169-183.

Desrosières, A. (2008). *Gouverner par les nombres* Paris: Presses de l'école des mines.

European Commission (2011). *Innovation Union Competitiveness Report* Brussels: .

European Court of Auditors (2007). *Evaluating the EU Research and Technological Development (RTD) framework programmes* Brussels: European Commission.

Feller, I. (2007). Mapping the frontiers of evaluation of public-sector R&D programs. *Science and Public Policy, 34(10)*, 681-690.

Georghiou, L. & Larédo, P. (2006). *Evaluation of Publicly Funded Research – Recent Trends and Perspectives* Paris: OECD DTI/STP(2006)7.

Glänzel, W. & Schubert, A. (2005). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel & U. Schmoch(Eds.) *Handbook of Quantitative Science and Technology Research* (pp. 257-276). Dordrecht: Kluwer Academic Publications.

Godin, B. (2005). *Measurement and Statistics on Science and Technology* London: Routledge.

Grupp, H., Schmoch, U. & Kuntze, U. (1991). Patents as potential indicators of the utility of EC Research Programmes. *Scientometrics, 21(3)*, 417-445.

Grupp, H. (2000). Indicator-assisted evaluation of R&D programmes: possibilities, state of the art and case studies. *Research Evaluation, 8(2)*, 87-99.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth Generation Evaluation* Newbury Park et al.: Sage.

Gulbrandsen, M. & Slipersaeter, S. (2007). The third mission and the entrepreneurial university model. In A. Bonaccorsi & C. Daraio(Eds.) *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe* (pp. 112-143). Cheltenham: Edwar Elgar.

Guston, D. H. (1999). Stabilizing the Boundary between US Politics and Science:: The Rôle of the Office of Technology Transfer as a Boundary Organization. *Social Studies of Science, 29(1)*, 87-111.

Guy, K. & Nauwelaars, C. (2003). *Benchmarking STI policies in Europe: In search of good practice* Seville: Institute for Prospective Technological Studies.

HEFCE (Higher Education Funding Council for England) (2007). *Research Excellence Framework. Consultation on the Assessment and Funding of higher education research post 2009* London: HEFCE.

Heller-Schuh, B., Barber, M., Henriques, L., Paier, M., Pontikakis, D., Scherngell, T., Veltri, G. A. & Weber, M. (2011). *Analysis of Networks in European Framework Programmes (1984-2006)* Luxembourg: Publications Office of the European Union.

IPTS (2002). *RTD Evaluation Toolbox* Brussels: Institute for Prospective Technological Studies, EUR 20382N.

Joly, P. B. & Mangematin, V. (1996). Profile of public laboratories, industrial partnerships and organisation of R&D: the dynamics of industrial relationships in a large research organization. *Research Policy, 25*, 901-922.

Kanninen, S. & Lemola, T. (2006). *Methods for Evaluating the Impact of Basic Research Funding* Helsinki: Academy of Finland.

Klerkx, L. & Leeuwis, C. (2008). Delegation of Authority in Research Funding to Networks: Experience with a Multiple Goal Boundary Organization. *Science and Public Policy, 35(3)*, 183-196.

Kuhlmann, S. (2003). Evaluation of research and innovation policies: a discussion of trends with examples from Germany. *Journal of Technology Management, 26(2-4)*, 131-149.

Kuhlmann, S. (1998). Moderation of Policy-making? Science and Technology Policy Evaluation beyond Impact measurement: the Case of Germany. *Evaluation, 4(2)*, 130-148.

Kuhlmann, S., Boekholt, P., Georghiou, L., Guy, K., Héraud, J., Larédo, P., Lemola, T., Loveridge, D., Luukkonen, T., Polt, W., Rip, A., Sanz Menéndez, L. & Smits, R. (1999). *Improving distributed intelligence in complex innovation systems* Karlsruhe: Frauenhofer Institute - Systems and Innovation Research.

Larédo, P. & Mustar (2000). Laboratory activity profiles: An exploratory approach. *Scientometrics, 47(3)*, 515-539.

Latour, B. & Woolgar, S. (1979). *Laboratory Life. The construction of scientific facts* New Jersey: Princeton University Press.

Lepori, B. (2011). Coordination modes in public funding systems. *Research Policy, 40(3)*, 355-367.

Lepori, B., Barré & Filliatreau (2008). New perspectives and challenges for the design and production of S&T indicators. *Research Evaluation, 17(1)*, 33-44.

Lepori, B., Dinges, M., Reale, E., Slipersaeter, S., Theves, J. & Van den Besselaar, P. (2007). Comparing the evolution of national research policies: what patterns of change? *Science and Public Policy, 34(6)*, 372-388.

Lissoni, F., Llerena, P., McKelvey, M. & Bulat, S. (2008). Academic Patenting in Europe. New Evidence from the KEINS database. *Research Evaluation, 17(2)*, 87-102.

Luukkonen, T. (1998). The difficulties in assessing the impact of EU Framework Programs. *Research Policy, 27(6)*, 599-610.

Martin, B. R. & Irvine, J. (1983). Assessing Basic Research: some partial indicators of scientific progress in radio astronomy. *Research Policy, 12(2)*, 61-90.

Merkx, F. & Van den Besselaar, P. (2008). Positioning indicators for cross-disciplinary challenges: the Dutch coastal defense research case. *Research Evaluation, 17(1)*, 4-16.

Moed, H. F., Glänzel, W. & Schmoch, U. (2004). *Handbook of Quantitative Science and Technology Research* Dordrecht: Kluwer Academic Publishers.

Mogoutov, A. & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *36 (6)*, 893-903.

Molas-Gallart, J. (2011). Tracing 'productive interactions' to identify social impacts: an example from the social sciences. *Research Evaluation, 20(3)*, 219-226.

Molas-Gallart, J. & Davies, A. (2005). Toward Theory-Led Evaluation - the Experience of European Science, Technology and Innovation Policies. *American Journal of Evaluation, 20(10)*, 1-18.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics, 66(1)*, 81-100.

OECD (2006). *Government R&D Funding and Company Behaviour. Measuring Behavioural Additionality* .

OECD (2005). *Guidelines for Collecting and Interpreting Innovation Data — The Oslo Manual* Paris: OECD.

OECD (2002). *Frascati Manual. Proposed Standard Practice for Surveys on Research and Experimental Development* .

OECD (1995). *The Measurement of Human Resources devoted to S&T - Canberra manual* Paris: OECD.

Porter, T. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* Princeton: Princeton University Press.

Proneos, OST, NIFU-Step & CERIS-CNR (2009). *Tools and Indicators for the Evaluation and Monitoring of European Research Framework Programmes* Brussels: European Commission.

Reale, E., Inzelt, A., Lepori, B. & van den Besselaar, P. (2011). *Indicators for the evaluation of the Internationalization of Governement Funding Agencies: results from an exploratory and participatory project at European level* .

Roessner, D. (2000). Quantitative and qualitative methods and measures in the evaluation of research. *Research Evaluation, 8(2)*, 125-132.

Rogers, J. & Jordan, G. (2011). Introduction to a special section on new approaches to predicting and tracing R&D program effects for policy learning. *Research Evaluation, 20(4)*, 263-266.

Ruegg, R. & Feller, I. (2003). *A Toolkit for Evaluating Public R&D Investment: Findings from ATP's First Decade* NIST GCR 03-857.

Scherngell, T. & Barber, M. (2011). Distinct spatial characteristics of industrial and public research collaborations: evidence from the fifth EU Framework Programme. *The Annals of Regional Science, 46(2)*, 247-266.

Spaapen, J. & van Drooge, L. (2011). Introducing 'productive interactions' in social impact evaluation. *Research Evaluation, 20(3)*, 211-218.

Stake, R. E. & Abma, T., (2005).
Responsive evaluation. In S. Mathison(Ed.) *Encyclopaedia of Evaluation* (pp. 376-379). Thousand Oaks: Sage.

Stame, N. (2004). Theory-based Evaluation and Types of Complexity. *Evaluation, 10(1)*, 58-76.

Technopolis & OST (2009). *Bibliometric profiling of Framework Programme participants* Brussels: .

Technopolis & OST Paris (2004). *Future priorities for Community research based on bibliometric analysis of publication activity for the Five-year Assessment (1999-2003) of Community research activities* Brussels: EPEC.

Thoenig, J. -. (2000). Evaluation as usable knowledge for public management reforms. *6*, 217-229.

Tijssen, R. (2001). Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows. *Research Policy, 30(1)*, 35-54.

Trochim, W. M., Marcus, S. E., Masse, L. C., Moser, R. P. & Weld, P. C. (2008). The evaluation of large research initiatives - A participatory integrative mixed methods approach. *American Journal of Evaluation, 29*, 8-28.

van Raan, A. F. J. (2004). Measuring science. In H. F. Moed, W. Glänzel & U. Schmoch(Eds.) *Handbook of Quantitative Science and Technology Research* (pp. 19-50). Dordrecht: Kluwer Academic Publishers.

Weiss, C. (2004). Theory-based Evaluation: Past, present, future. *New Directions for Evaluation, (76)*, 41-55.

Whitley, R. & Glaser, J. (2007). *The changing governance of the sciences. The advent of research evaluation systems* Dordrecht: Springer.