INNO-Appraisal

Understanding Evaluation of Innovation Policy in Europe

Executive Summary

February 2010
Executive Summary

Manchester Institute of Innovation Research (formerly PREST),
MBS, The University of Manchester
Prof. Jakob Edler (Project Coordinator)
Dr Paul Cunningham
Abdullah Gök
Dr John Rigby

Atlantis Consulting
Effie Amanatidou
Ioanna Garefi

Fraunhofer Institute
Systems and Innovation Research

ISI-Fraunhofer
Dr Susanne Bührer
Dr Stephanie Daimer

Joanneum Research
Michael Dinges
Dr Martin Berger
Julia Schmidmayer

Wise Guys Ltd.
Wise Guys Ltd.
Ken Guy

1 Wise Guys Ltd. formally left the project on May 15, 2009, though Ken Guy continued to contribute to the project in his capacity as a visiting scientist at JRC-IPTS, Seville.
This project breaks new ground. For the first time, it takes stock, categorises, and analyses evaluation practice in innovation policy in EU Member States. The project ran from January 2007 to January 2010 and would have not been possible without the help and support of many people.

During all of this time, a steering committee has overseen or followed the work, reflected on its various steps and come together to two key events. The advice given by the members of that committee has been extremely valuable. The project team would like to thank the policy makers Ulrike Blankenfeld (DE), Mark Beatson (UK), Luisa Henriques (ES) Nick Constantopoulos (GR), Rupert Pichler (AT), Jari Romanainien (FI), as well as our colleagues Aleardo Fulrani and Bart Kamp (INNOVA Europe), Phil Shapira (MBS, Manchester) and Anthony Arundel (MERIT) very much for their advice and for their valuable time. Without this advice, the reassurance and constructive critique and the ideas for new avenues the study team would have felt much less secure in pushing ahead with such a novel methodology.

We also would like to thank DG Entreprise for the opportunity to do such rewarding work. An especially heartfelt thank you goes to Alberto Licciardello, who has done a tremendous job as project officer to help to keep it all together and focused, to support in times of need, to coordinate with other services, to inspire the analysis and to disseminate our activities to colleagues inside and outside of the Commission. Finally, the project would like to thank all the participants at the INNO-Appraisal workshop in September 2009; the feedback gained at this event has been tremendously important and encouraging in the final stages of this project. Finally, our thanks goes to the numerous policy makers across Europe who have participated in this study, by filling in templates about the evaluations they have commissioned and by agreeing to be interviewed and otherwise give feedback. Without this openness and responsiveness, the study would simply have been impossible.

On behalf of the INNO-Appraisal team

Jakob Edler
The aim of the INNO-Appraisal project was to contribute to a better understanding of how evaluation is currently used in innovation policy in Europe, and how evaluation contributes to policy making. INNO-Appraisal was the first systematic attempt to provide an overview of evaluation practice in Europe. By doing so, it sought to achieve a second, equally important aim: i.e. to render evaluation practice accessible to the policy and evaluation community. A third aim was to contribute to a better-informed evaluation discourse across a better-networked evaluation community in Europe.

To achieve these aims, the project spent three years taking stock of and assessing evaluations in the area of innovation policy across Europe. It applied a novel and complex approach, combining qualitative in-depth analysis (case studies) and sophisticated quantitative analysis on the basis of a new form of data collection. The basis for the evaluation report collection was the EU innovation policy database Trendchart over the period 2002 to 2007. The project designed and made use of a web-based template to allow a systematic characterisation of all selected evaluation reports. It then interacted with policy makers in order to verify and amend these characterisations. The template data was then used to conduct a statistical analysis of the whole sample and further analyses of subsamples relating to specific questions and case studies. To make these evaluation reports accessible, a repository of evaluation reports was created and placed on the INNO-Appraisal webpage. This repository allows interested parties to search for and download evaluation reports. It was also designed to allow a keyword search using the categorisation scheme on which the analytical template for each report was based. Thus, policy makers can now perform specific queries tailored to their specific needs, e.g. searches for examples of the application of particular methods, the coverage of certain topics or the evaluation of similar types of programme.

Descriptions of the approach adopted by the project, its various interim results and the repository itself have also been widely disseminated throughout the PRO—INNO® community and the wider policy and analyst community in innovation policy in Europe.

Thus, the major contributions of the project to the evaluation community in Europe and evaluation discourse in general are:

1. An analysis of evaluation practice with some in-depth topic-oriented and country case studies (this report),
2. the repository on the INNO-Appraisal webpage, with all its various search and download functionalities and its legacy role as a stockpile of evaluation reports and activities (http://www.proinno-europe.eu/appraisal).

In conjunction with earlier reports, presentations and interactions detailing interim results, this report (and subsequent outputs) and the repository itself should contribute to an improved policy discourse in the EU and beyond. It needs to be stressed, however, that the report does not constitute another ‘evaluation manual’. Rather, it is analytical in nature and provides a service to the Community by making evaluation in Europe more ‘tangible’.

This executive summary encapsulates the major findings of the analytical part of the project in some detail to reflect the depth and breadth of the analysis and to avoid undue simplifications.
1. The characteristics of evaluations in innovation policy in Europe

1.1. In a nutshell: basic characteristics of evaluation practice in innovation policy in Europe

The INNO-Appraisal database contains evaluation of a whole range of different policy measures that are covered in the Trendchart Database between 2002 and 2007. Reflecting the innovation practice across Europe, its majority of evaluations are concerned with direct financial support for innovation activities, and two thirds of the underlying sample measures in the database are geared towards involving Universities and public research institutes. The database of evaluations covers all European countries, with an interesting bias towards Austria which has an exceptionally high number of innovation policy measures reported in Trendchart and an extensive evaluation activity. The repository, as basis for the analysis, cannot claim to cover all innovation policy evaluation in all countries to the same degree, as countries represent their activities differently in the Trendchart database, the basis of the analysis. For example, Finland and Sweden are underreported in the database as many of their evaluations are undertaken within the programme portfolio of large agencies so that many individual measures are not flagged up in the Trendchart database. To get an understanding of the meaning of country contexts, however, later sections deliver in-depth country cases of Austria, Germany, UK and Mediterranean countries. The repository also covers extensively evaluation of structural fund measures, as slightly more than 20% of all evaluations are performed in the context of structural funds. The total number of evaluation reports is 242, of which 216 could be meaningfully analysed by the project team (and thus used for the statistical analysis presented), and 146 were amended and verified by policy makers (used for specific, judgemental and policy related parts of the statistical analysis). The number of publishable evaluation reports in the project repository is 173.²

Commissioning and design: Evaluation is found to be an integral part of innovation policy, as roughly 50% of the measures that are evaluated have a pre-determined budget for evaluation and two thirds are foreseen and planned in the measure design. More than 90% of evaluations are sponsored by the programme owners themselves, only a minority are jointly sponsored with other bodies or entirely external (10%). Almost half of the evaluations follow an open tender procedure, one fifth are done through closed tender, one fifth are performed by external evaluators without a tender and 15% are done internally. For those evaluations which have a tender, a large majority clearly specified the objectives, whilst at the same time, two thirds of the tender documents left the choice of methods to the evaluators.

Timing: More than 40 % of the database is interim evaluations. This bias against ex post (30%), however, stems partly from the selection method focusing on live Trendchart policies within a certain period of time. The database contains both formative (33%) and summative (21) evaluations, while the majority combines both summative and formative aspects.

Topics: The topics covered in evaluations are obviously broad. In very general terms, effectiveness and consistency appear to be slightly more important than programme efficiency issues, while the in-depth look at project efficiency is much less common (below 50%). We also find a certain clustering of topics covered. Two thirds of all evaluations cover at least one form of additionality

² For a series of methodological and database specific reasons one cannot give a statistical data as for the share of policy measures that are evaluated within Trendchart in the period covered.
(input, output and behavioural), and many of those evaluations tend to include the project level in order to understand those additionalities. Gender (24%) and minority (7%) are least common. In terms of impact, technological and economical are most important, and environmental impacts (still) least important (28%).

**Methodology and data sources:** In terms of methodology, we find a whole range of methods applied, however, some general strong trends are obvious. Descriptive statistics are the most common approach, applied by more than three quarters of all evaluations, while case studies – to understand contexts and developments over time – are performed only by 41%. More sophisticated, quantitative approaches are used even more selectively, e.g. 23% perform econometric analysis, 17% network analysis. Interestingly, 80% claim to use monitoring data and 70% to use existing surveys and databases as a basis for the analysis. However, it appears that this kind of data is insufficient to be used for specific evaluation questions such as networking or behavioural additionality. The most important pro-active data collection is done through interviews and participant surveys. Technometric analysis in innovation policy plays no significant role at all (2%), it appears that for the analysis of technological substance in projects peers are used (20%).

**Quality:** As for overall quality of evaluations, the database shows very mixed results along nine different quality aspects. For a general picture a simple binary quality index has been constructed. All evaluations that score more than 3 on a Likert scale (1 being very low, 5 being very high) in each of four selected quality variables are defined as being of high quality. 61% of the evaluations show an overall positive quality index. This means that almost 40% of the evaluations have serious quality problems in at least one key quality dimensions. This finding is confirmed through an auto-correlation analysis: Many evaluations are either good in a whole set of quality criteria or perform rather badly across the board.

**The policy use of evaluations:** While almost all evaluations are targeted towards policy makers and programme management, only 50% of the evaluations are targeted towards the users of the programme and less than one third to the general public. Evaluations are obviously not extensively used to mobilise the community, policy makers themselves rate the breadth and depth of actual discussion about evaluation results only as moderate.

Most evaluations, as is to be expected, contain recommendations for policy and programme management, only a minority of evaluations is purely analytical. The usefulness of the recommendations for various aspects of policy learning and improvement that were tested is moderate and appears to have room for improvement. In principle, evaluations are not linked with major, radical consequences, those appear to be the result of more general policy considerations. However, they are important for minor re-design of measures or their prolongation and extension. In 17% of all cases they are also used to improve other or future policy measures.

### 1.2. Determinants of evaluation practice, quality and consequences

There is a certain degree of convergence of evaluation practice across different policy measure. We find surprisingly little variation between different policy measures as regards a whole range of evaluation characteristics, such as tender procedures, internal vs. external evaluators, coverage of topics and impacts and even use of some of the data collection approaches and methods and even targeted audiences. It shows that other factors, such as organisational and country specific
traditions, topics to be covered and general practices dominate the design and implementation of evaluations to a large extent, not so much the evaluation object – the policy measure – itself.

However, the type of measure makes some difference as for evaluation design and implementation. Certain specific types of programmes show a specific application of tailored methods and data collection approaches (e.g. network analysis and case study approaches for networking and cluster programmes). We also find variation in the use and dissemination of evaluation between policy measures, with - for example – complex networking programmes targeting beneficiaries much more often as those measures are complex and need explanation and formation. Furthermore, evaluations for direct financial support measures and cluster, technology transfer and networking measures are more likely to be perceived as being of good quality, while evaluations for softer measures such as management support measures or diffusion measures are of lower quality. In addition, there seems to be a poorly developed evaluation practice for diffusion measures, which – in addition – do not take societal and environmental impacts into account as broadly as expected, and that are perceived to be of less usefulness to policy makers.

Evaluations are often influenced by external sponsors of the policy measures. While they do not impose methods they introduce a bias towards social and environmental impacts and gender and minority issues. The external sponsors, it seems, are one major reason behind a certain grouping of evaluations around topics that we observe, some being more concerned with quantitative, hard economic and technological outputs, and others interested in social, environmental impacts etc.

Evaluators in innovation policy appear to apply a form follows function approach; they tailor their approaches according to the need for topics and impacts to be covered. For example, evaluations interested in strategy development and policy issues more generally also look at consistency and vastly use interviews and other qualitative methods. Evaluations more concerned with effectiveness rely on (often simple) statistical analysis and data, and the use of peers, although limited, is strongly linked to quality of output. Those evaluations more concerned with efficiency and project level issues, in turn, tend to look for different kinds of additionality and rely on surveys, interviews and, less broadly, though, on case studies. Further, formative and ex ante evaluations tend to analyse consistency issues more broadly than other evaluations (i.e. to assess and re-adjust the overall match), and they do so be using slightly more qualitative methods.

A deeper look into the determinants of quality assessments reveals that policy makers see room for improvement as regards the coverage of the broader context, the application of advanced quantitative and some qualitative methods and the documentation of information sources. In contrast, evaluations covering technological and scientific impact and those using survey methods and peer review are perceived as being of a higher quality. Summative evaluations appear to be perceived as being of higher quality than formative evaluations, and indeed they are more widely discussed within government than formative ones. Formative evaluations, it seems, are a tool for improvement for the programme owners and beneficiaries, while the messages of summative evaluations are used for wider discourse and justification.

Interestingly, quality does not differ between evaluations that are undertaken by external evaluators and those performed internally. Equally, evaluations are not perceived to be of higher quality if they are in-built in policy measures from the start and have a dedicated budget within the policy measure. However, one important finding is that quality is lower for evaluations that are
commissioned by external sponsors or policy bodies. In contrast, open tenders yield evaluations of better quality.

**Quality**, finally, makes a difference when it comes to the dissemination and exploitation of evaluations. The better an evaluation, the more likely it is discussed within and outside government. In addition, evaluations that ex ante are targeted to the wider public and policy analysts (and not only to the programme management) are also of higher quality.

The analysis also revealed that evaluations have a limited set of consequences. Radical consequences (termination of programmes) are very rarely a result of an evaluation, but rather they appear to be the consequence of principle policy decisions. In contrast evaluations lead to minor re-design of measures or learning for other measures and, most often, to prolongation and extension. The latter is highly correlated with simple methods, it thus appears that clarity and simplicity in the data and methods is part of a confirmation and incremental approval exercise. In addition, those evaluations which are intensively discussed within and outside government are those that are more likely to lead to consequences. Finally, quality also is important for evaluation consequences; evaluations of higher quality more often tend to lead to consequences (especially prolongation). The quality aspects most strongly linked to the likelihood for consequences out of evaluations are the extent to which evaluation methods satisfy the Terms of Reference and the purpose of the evaluation.

In a final analytical step the general statistical analysis explored clusters of evaluations. Two clusters emerged. One cluster of evaluations is more populated by ex ante evaluations and is concerned with programme efficiency issues and, by nature, more often based on qualitative methods. The second cluster appears to be more ex post and interim, being broader in its coverage and more concerned with different forms of outcome/impact, thereby mobilising more quantitative approaches and oriented towards the policy community rather than the beneficiaries. This cluster of evaluations is more often used for decision about prolongation or re-design of measures.

2. **In-depth analysis of selected evaluation issues**

Four themes of evaluation have been identified as being of specific importance to stakeholders and the evaluation community have been studied in considerable depth; usefulness, measurement of impact, behavioural additionality and structural fund evaluation.

2.1. **Usefulness of evaluation**

The analysis of the usefulness (or utility) of evaluations sets the broader context of policy interventions within a policy mix, and the accompanying need for policy makers to be able to judge the effectiveness and efficiency of their interventions through the use of a range of governance tools, including appraisal, monitoring and evaluation. It is clear, from the policy mix concept, that the information gained from these tools should not be restricted to the subject of the assessment, but should also be relevant to the design and operation of contemporaneous or subsequent policy instruments: such requirements define the issue of usefulness.

The report then discusses what is meant by usefulness and utility in the context of the evaluation of innovation support measures. **Three major purposes** for evaluation are identified: operational
learning, policy feedback and system impact. Overall, it is suggested that, to be useful, evaluations must provide information on: the effectiveness of design, the effectiveness of management, the effectiveness of implementation, the effectiveness of the evaluation itself, the achievement of objectives, and the broader impacts of the instrument. However, it is recognised that usefulness may also be impacted by other factors such as audiences and sponsor demands.

A number of factors are then examined whereby the utility of evaluations may be increased. Possible routes include increasing the rigour (and hence ‘quality’) of an evaluation, obtaining the compliance and trust of stakeholders, improving the transparency of methodologies (assuming an informed audience of policy makers is present), and the use of clear and measurable objectives. The incorporation of evaluation into the overall policy cycle is seen as a clear route to improving the usefulness of its outcomes.

The chapter next deals with the approaches employed in the analysis of the survey results to determine the extent of usefulness of the evaluations reviewed. Two lines of analysis were followed: looking for evidence of utility provided by the responses and testing of hypothesised links between utility and other database variables. As the questionnaire did not specifically pose a direct question on the usefulness of the evaluation (which would have prompted highly subjective responses unsuitable for quantitative analysis), it was necessary to develop a proxy for usefulness based on the extent to which the evaluation report’s recommendations had been useful (a point addressed by specific questions in the questionnaire template). This proxy indicator (for overall usefulness) could be broken down into internal utility (relating to changes to the programme under appraisal) and external utility (relating to changes to contemporaneous or subsequent programmes).

The analysis then examined a number of the database variables for links with usefulness. The main points to emerge were:

- 84% of evaluations examined had contained recommendations, with an almost equal balance between internal recommendations (relevant to the subject programme) and external recommendations (relevant to future programmes or to broader policy formulation).
- Evaluations addressing internal aspects of the programme had a slightly higher usefulness than those addressing external aspects.
- Significant positive correlations with at least one aspect of usefulness were identified for:
  - The use of an open tendering process when commissioning and evaluation
  - The use of external evaluators
  - The timing of the evaluation (ex ante, interim, ex post, etc.)
  - Summative over formative evaluations
  - Non-Structural Fund evaluations (i.e. a negative correlation between Structural Fund evaluations and utility)
  - Non-portfolio type evaluations (i.e. a negative correlation between portfolio type evaluations and utility)
  - Non-conditional evaluations (i.e. a negative correlation between conditional evaluations and utility)
  - Evaluations that examined the topics of goal attainment and effectiveness and policy/strategy development
Executive Summary

- Evaluations that examined scientific impact and technological impact on the participants and beyond
- Evaluations that employed case study analysis; participant surveys; interviews; focus groups/workshops and meetings; peer review
- Evaluations that resulted in a minor redesign or expansion/prolongation of the measure
- Evaluations sponsored by programme managers, other government departments or other public bodies
- Evaluations not conducted primarily for auditors/financial authorities
- Evaluations whose reports were published in English
- Certain dimensions relating to the quality of the evaluation

- **Negative correlations** with at least one aspect of usefulness were observed for:
  - Evaluations that examined input additionality and environmental impacts
  - Evaluations that employed input/output analyses; context analysis; group comparison approaches; cost/benefit approaches; existing surveys and databases

- **No significant correlations** with any aspect of usefulness were detected for:
  - Evaluations planned during the design of the measure
  - Presence of a dedicated budget for the evaluation
  - Evaluations conducted primarily for policymakers (government officials) and programme management
  - Evaluations that examined outputs, outcomes and impacts; quality of outputs; value for money; programme/project implementation efficiency
  - Evaluations that employed monitoring data
  - Evaluations that had wider levels of availability
  - Evaluations where a major redesign of the measure resulted

- External utility was more highly rated in Germany and the Netherlands, whilst internal usefulness was more highly rated in Greece, Sweden and the UK

- The evaluations of measures for science-industry cooperation were significantly more useful across all categories of usefulness. Evaluations of measures aimed at the creation of start-ups and spin-offs were also significantly useful (external and overall).

Whilst a number of the statistically significant associations between usefulness and the survey variables were anticipated, it is harder to explain some of the negative correlations or where no correlations were detected. Several of the latter might be explained by the relatively low number of cases available within the analysis, whilst the prevalence of Structural Fund evaluations within the sample could also provide an explanation.

In conclusion, the results of the analyses present a mixed picture, confirming some expectations yet failing to confirm or even refuting other expectations. As with most research endeavours, it is clear that further investigations are required into the aspect of usefulness and it is hoped that this study offers a valuable starting point. Nevertheless, the results do tend to support the overall conclusion (which is also based on the direct input of policymakers in the field) that usefulness is a highly subjective and context specific issue and that, as a broad rule of thumb, an evaluation may be considered useful if it delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers evidence as for value for money. Usefulness can be
defined as the degree to which there is feedback on policy and if the evaluation process delivered some degree of policy learning.

2.2. Measuring impact

While there is extensive academic debate about an ideal-type setting for impact measurement in evaluations, it is a quite different matter how impact assessments are performed in reality. Most often, impact assessment is rather limited and simplistic in its approach within the reality of service contracts for programme owners, most probably including budget restrictions, specific “customer” needs, and tough schedules. One in-depth analysis of the INNO-Appraisal database has looked systematically at the application of impact measurement in evaluation of innovation policy. It explored if there is some systematic use of methods, i.e. whether there are certain sets of methods which are employed for specific policy measures and in specific contexts, and whether, and also to what extent evaluation studies of policy programmes have an impact on future innovation policy.

In sum, the quantitative analysis of the database shows a number of interesting results: In general,

- Impact assessment is a central function of evaluation studies: A large number of studies across Europe claim to do impact assessment, currently most important are economic impacts.
- Impact assessments appear to be central and wide-spread across Europe
- Impact studies of structural fund evaluations differ significantly from impact studies of national innovation programmes.

As for specific types of impact, we find that

- Typically the use of a very broad definition of impact assessment, including all types of effects
- Assessment of economic impact is most dominant, other impact types of importance are technological and societal impacts (not scientific and environmental impacts)
- The assessment of new impact types (apart from economic or technological) is still rather uncommon. Societal impacts are often covered with an estimation of new jobs having been created, but other topics, such as gender impacts are quite rare.
- A high number of evaluations claim to assess indirect impacts, i.e. spill-over effects beyond the participants of a programme. This is, given the methodological difficulties for assessing economic or societal impacts, a surprising result. This result seems to reflect the demand for results on these spill-over effects.
- Additionality concepts are well established beyond the UK. They are employed by half of the evaluations in the sample. This is also true for behavioral additionality which has obviously become an integral part of the idea of additionality.
- Structural fund evaluations more often cover social and environmental impacts.

Methods used

- Almost the whole toolbox of possible methods is employed for impact assessment, including elaborate methods such as a control group approach.
- Most of the impact assessments are qualitative and part of broader evaluation studies.
There are only few quantitative impact assessments using elaborated quasi-experimental designs like control-group approaches.

Impact assessment is typically not a mere econometric exercise, but *often used in a contextually sensitive way*.

### Policy Cycle

- Impact assessment is not a clearly retrospective element of evaluation. Often, it is also used in the form of ex-ante impact assessment and in accompanying evaluations.
- Evaluations which include impact assessments, in particular the assessment of societal impacts, are more often used for external communication. Experts confirm that impact assessment is in particular important for legitimizing the political interventions.
- If impact assessments are included into evaluations this leads to higher quality scores.
- With respect to usefulness, evaluations of (single) national programmes seem to be more useful for policy makers than structural fund evaluations.

The analysis and the interviews indicate a set of clear recommendations. Most important issues from the perspective of the policymaker are:

1. Impact assessments are an important part of evaluations, but *should not be isolated*. Ideally, impact assessment is *integrated into a broader, more holistic evaluation framework* (e.g. covering context analysis systematically), only then can it fully be understood.

2. Evaluators have responded to the demand for quantitative results and employ a variety of (elaborate) methods to achieve them. However, in most cases it seems that *the combination of qualitative and quantitative analysis* can cope more adequately with impact assessments, as many impacts are not quantifiable at all.

3. Many pitfalls of impact studies can be avoided by a *constant communication between policymakers and evaluators* during the process of evaluation. This leads to transparency for the whole evaluation process in order to realize learning and to cope with methodological challenges.

4. As impact assessments clearly pursue the two purposes of learning and legitimation, two types of recommendations might be considered as a result of impact evaluations: those designed for *policy improvement* implemented by the programme owners / managers and those directed to higher levels, which serve the legitimation aspect.

For the future, it is useful to consider further impact dimensions to a greater extent than in the past. Additionally, as we expect more mission oriented policy programmes where other topics like sustainability, customer needs and the structural / regional development might become more important. Thus, impact assessment will have to be *broadened considerably in the future*. Further, with examination of the demographic challenges and shortages of a skilled workforce in most European countries, the issue to integrate larger parts of society to the research sector will become even more relevant than in the past and therefore impact assessments should also address gender and minority issues in more detail. Finally, the still prominent aspect of Behavioural Additionality in Innovation Programmes (e.g. innovation management, risk aversion) will remain important.
For impact assessment, this all means that it will become **even more demanding** to measure the intended effects – at least quantitatively. Given that non-economic impacts will gain more and more in importance this would mean that new sets of criteria and indicators will have to be defined, and most likely many of these indicators will be of a qualitative nature. More public support for **experimental evaluation designs** (including meta-evaluations at national as well as European level) could help to identify the most promising ways to identify new impact types.

However, given the multiplication of goals and increasing pressure as to economic effects impact assessment, the requirement to establish the relevance and rank different impact dimensions will become all the more necessary. Equally, policy must reflect if, and to what extent the large set of impact dimensions can really be achieved by one single measure instrument. The programme objectives have to correspond with an appropriate mix of policy instruments and have the right balance between direct and indirect funding. Additionally, policy design has to be very aware about the prerequisites for (behavioural and system) change which cannot entirely be influenced by singular measures. By definition, impact assessment can only be one, even if essential, part of evaluation to support those policies.

**2.3. Behavioural additionality**

Behavioural Additionality is still a rather novel, but already a **key topic for evaluations**. The concept has **enlarged** our thinking about the **effects of innovation policy** to include, more systemically, **learning** as a key outcome in itself, enabling further, and broader, and more sustainable innovation. Behavioural evaluation is a case of reflexive intelligence, whereby working on understanding the concept and applying it to innovation policy itself co-evolves with innovation policy concepts that take learning into account much more profoundly. Evaluation practice and conceptualisation on the one hand and innovation policy development on the other hand have re-enforced each other. The empirical analysis on behavioural additionality in this study rests on three pillars, a statistical analysis, a text analysis of evaluation reports and a set of interview based case studies of evaluations.

In the **academic literature**, the term is understood in at least **four different conceptualisations of behavioural additionality**, namely, i) as an extension of input additionality, ii) as change in the non-persistent behaviour related to R&D and innovation activities, iii) as change in the persistent behaviour related to R&D and innovation activities and iv) as change in the general conduct of the firm with substantial reference to the building blocks of behaviour.

Against this background, a text analysis of 33 selected evaluations demonstrated that the diversity of understandings is reflected as well in **evaluation practice**, where we also find **at least four different understandings** of the concept. They are distributed rather evenly in the sample – and thus there is yet no dominant understanding established just like the case in the scholarly literature. These understandings differ in the their conceptual outreach, ranging from collaboration (non-persistent) in R&D and innovation only – the most narrow type – to persistent change in management practices more broadly, beyond R&D and innovation – the broadest type. The types overlap, but not entirely match four ideal types as defined in the vast literature on the concept.

The analysis of the INNO-Appraisal database aims to show if and how evaluations differ that apply the concept from those that do not. For the first time this allows us to get a systematic picture of the
nature of behavioural additionality in practice. The core results are as follows.

The data analysis shows that behavioural additionality is a well established concept in evaluations, **50% of all reports in the database employ it**, explicitly or implicitly. The concept is more often used for networking and technology transfer concepts, which is consistent with the need for learning, networking and cooperation in those programmes. The behavioural additionality concept is most often used in conjunction with input and/or output additionality. It appears to be more important in evaluations that are also concerned with **project level evaluations**, not only programme level, which again is consistent with the basic idea of understanding the micro level in order to understand the macro effect. The concept is **less common in portfolio and structural fund evaluations** as those often do not look at the project level.

While there is no difference between evaluations that are sponsored by the programme owners themselves or by other bodies, we observe that the concept is slightly less often applied in internal evaluations. The application **needs specific expertise** and in-depth qualitative approaches which seem to be best conducted by external evaluators. However, this does not imply that evaluators are more keen to apply it than policy makers, since the concept is more often applied in those evaluations in the database that specify the methodology in the terms of reference – and thus express a **clear demand for behavioural additionality** approaches. Our in-depth case studies indeed confirm that both evaluators and policy makers can be the source for the application of the concept, it is not entirely evaluator driven.

Interestingly, and neglecting its full potential, behavioural additionality is not as common in accompanying evaluations as one would assume given the focus on interaction and learning and the need to re-adjust programme and implementation should learning effects not be observed in real time. The concept is used in formative evaluations, but not as extensively as one would think. Similarly, evaluations that cover behavioural additionality are less likely to look at social and environmental impact, but **much more at scientific and technological impact** than the whole sample, while the concept is equally concerned with economic impact than the whole sample of evaluations.

As for methods, behavioural additionality evaluations are **more qualitative** and apply those methods with greater quality, however, the extent of case study analysis is not as broad as one would expect. Behavioural additionality evaluations also **use surveys more often**, while they cannot rely on **existing data or monitoring data**, pointing towards a need for adjusted monitoring.

Behavioural additionality evaluations are **broadly discussed** across government and beyond government, and they are **more often targeted towards the general public and towards users**. All this points to the **learning and mobilisation** potential of the concept. However, evaluations applying behavioural additionality are not perceived to be significantly more useful for changes in policies than other evaluations (although they perform slightly better in this regard). In terms of concrete consequences of the evaluations that apply behavioural additionality, the major difference to the general dataset is that the former lead **significantly more often to the extension of existing measures**. This again points to the underlying understanding of long term effects and the need for time in programmes that rely on the learning of actors.
The case studies finally confirm the variety of understandings and different application of the concept and the challenges the application of the concept faces. This is true both at the receiving end, with the programme owners, and at the performing end, with the analysts. The cases show that evaluators and policy makers alike are keen on understanding changes in behaviour better, but they also confirm that policy makers strongly demand a demonstration as to how the behavioural change translates into the intended innovation effect. However, many variables influence change in innovation, attribution remains a constant challenge and innovation effects often take considerable time to realise. Evaluations thus must clearly demonstrate the conceptual link between behavioural change and the innovation effect. They then must empirically grasp the change in behaviour and try to find robust indications that the conceptual link to innovation effects exists.

As yet, the applied methodologies do most often not fully capture behavioural additionality. The cases however show that it is possible to differentiate behavioural additionality and define building blocks of behaviour as well as chain of effects. This can be done in a mix of deductive and inductive approaches, with a focus on interaction with the beneficiaries. But there also is a delicate balance between exploring the concept to its full potential through all sorts of differentiation and methodologies on the one hand and pragmatic considerations and limits of absorptive capacity on the other hand. Thus, more experiments with sophisticated methodologies are called for. Those experiments should then enable us to define sets of meaningful, simplified methodologies that are more effective and efficient than the existing approaches, but do not overburden the process. To that end, there seems to be a huge potential in improving monitoring of programmes to use it for evaluations much more thoroughly.

Finally, the complexity of behavioural additionality asks for a strong interaction and communication between those commissioning the evaluation and the evaluators, since key concepts as to the link of behaviour changes to innovation must be shared between them and expectations clarified early on. Sophisticated methods alone do not guarantee the full benefits of the concept, their applications and the results must be intensively discussed among all stakeholders involved.

2.4. Evaluation in structural funds

The aim of this focus study is to examine if, and in what ways, the Structural Funds (SF) requirements and regulations related to evaluation influence the evaluation culture, institutional build up and good practice in evaluation in certain countries. It draws upon the results of the questionnaire survey carried as well as the examination of the uptake of SF regulations in three countries, Greece (a Southern European country) and two new Member States, Poland and Malta. The specific countries are examined as indicative examples of how SF evaluation related regulations and provisions are implemented and affect evaluation practices in their specific contexts.

The case study collection and analysis of data, information and stakeholder views is guided by the following hypotheses:

- SF requirements may lead to specific characteristics in delivery and practice of evaluation
- SF requirements may lead to higher quality evaluations
- High quality SF evaluations may have greater impact
- SF regulations demand high standards on structures and processes that inevitably need some institutional learning and structure building
SF regulations do seem to lead to specific characteristics in the delivery and practice of evaluation. They tend to be built in to the design phase of a programme/measure as they are a requirement in the SF implementation. They usually also meet the requirement to make the results publicly available through publication of the evaluation report. Recommendations mainly relate to the programme/measure being appraised in terms of design, management and implementation clearly reflecting the orientation of the SF evaluations.

SF requirements also seem to contribute to guiding the evaluation topics covered under the different evaluation types (ex-ante, interim, ex-post) as well as the data analysis methods used (but not the data collection methods). Yet, SF guidelines seem to more or less repeat what is suggested by international practice in evaluation and thus is also followed by non SF type evaluations. This might be the reason why no major differences exist when studying the results within the same evaluation type (ex-ante, or ex-post for example) across the two populations (SF and non SF).

SF requirements do not seem to lead to higher quality appraisals and even high quality SF evaluations do not lead to high impact in terms of usefulness of recommendations and dissemination of results. However, the suggestion to use independent (external) evaluators does seem to contribute to higher quality SF evaluations.

The country cases provide possible explanations for the survey results. The fragmentation among the key actors in the national innovation system in Greece, for example, and the fact that there is only typical abidance to SF regulations can explain why the results of SF evaluations are discussed with government and wider stakeholders only to a limited degree.

Abidance by the ‘letter rather than the essence of the law’, in combination with doubts about the suitability of the SF regulations leading to high impact evaluations can explain the limited usefulness of recommendations, as well as the fact that even high quality SF evaluations may not lead to high impacts in terms of usefulness and dissemination of results. The fact that SF regulations and quality standards are only suggested rather than imposed may explain why suggested quality criteria may not be applied in practice.

Finally the country cases show that while SF regulations have caused positive impacts in terms of capacity and structure building, they still fall short in improving institutional learning and establishing sound evaluation systems in the countries examined.

3. Country level analysis

3.1. Austria

Having been a laggard in terms of RTI (Research, Technology and Innovation) investments until the mid-nineties, both public and private entities have increased R&D investment efforts tremendously in the last decade. Austria has exceeded the average R&D intensity level of the EU-15 and the OECD countries. But not only RTI funding has increased: Austria has a large stock of innovation promotion measures at hand: Apart from generous bottom-up RTI project funding schemes, a remarkable number of thematic R&D programmes, structural programmes, and tax incentives exist. Despite good overall conditions there are a series of systemic challenges that still need to be addressed (e.g. poor performance of the Austrian higher education system, insufficient framework conditions as regards regulations, poor private and public funding for innovative start-ups and spin-offs).
During the catching-up process, RTI programmes were the most preferred way to address policy challenges. In this time, the use of evaluations increased dramatically. Evidence for the increased relevance of innovation policy evaluation is provided not only by evaluation counts, but by changes in the legal conditions for evaluations, measures to foster an evaluation culture, the transparency of evaluation results, and the high number of evaluation activities.

With 34 appraisal reports, Austria has the highest share of innovation appraisals in the INNO-Appraisal database. Some distinct features of these evaluations are presented.

The majority of appraisals are carried out mid-term during one point in the programme’s lifetime. Mainly, a supportive purpose is followed as policy makers respectively programme managers’ needs advise how to enhance programme implementation. Only a limited number of topics are addressed: Appraisals focus mainly on policy/strategy development, output counts, and consistency matters. Whereas behavioural additionality issues are rather prominent in Austria, input and output additionality issues as well as quality of outputs are only considered in a limited number of evaluations. Technological, economic, and socio-economic impact dimensions are missing by large, or only refer to programme participants.

Low cost data gathering and data analysis methods prevail (descriptive statistics, context analyses, interviews, and monitoring data). Most commonly a mixed methodological approach where quantitative and qualitative methodologies are combined is used.

Compared with the other countries in the dataset, we see a significantly lower coverage of input and output additionality issues, also the quality of outputs is widely neglected. Only a limited number of Austrian appraisals deal with impact at all: For every impact dimension, coverage is lower in Austria than in the other countries of the dataset. If impact dimensions are covered they focus on direct impact rather than on the participants and beyond.

Partly, the low coverage of impact dimensions and certain topics might be due to the formative purpose of most evaluations. Another reason for the discrepancies is the high coverage of Austrian appraisals in the database. Whereas in Austria almost the full range of appraisals conducted in the field of innovation policy is covered, it is more likely that only bigger evaluations are covered in the other countries; significant differences as regards the tender procedure point in this direction.

Despite the intermingled picture as regards evaluation topics used, the quality of evaluations is perceived to be high by respondents. Given the evaluation purpose, the methods used tend to be considered as appropriate. Especially, recommendations concerning changes to the management and implementation of RTI programmes were perceived to be useful. Forward-looking advice was regarded as helpful for the design and implementation of future policy measures.

Nevertheless, due to the high number of evaluative activities, an increasing evaluation fatigue can be witnessed. Criticism was raised, that mechanisms ensuring that the results of evaluations are fed back into policy formulation and implementation are missing. In this respect, more thoughts need to be spent on the concrete purpose of planned evaluation activities, and the role of evaluations for policy implementation.
3.2. Germany

Four major findings for innovation policies and the evaluation practice make Germany an interesting case to study and allow us to draw some recommendations on good evaluation practice.

First, innovation policies in Germany are focussing on high technologies, SMEs and the remaining special situation of the Eastern federal states. This is clearly reflected in the evaluated policy measures in the Inno Appraisal database.

Second, the institutional setup at the federal level provides for quite a systematic approach to evaluation. Almost all programmes are being evaluated. In particular the Ministry of Economics regularly foresees evaluations, when planning new programmes. Open tender procedures and the commission of evaluations to external evaluators are standard. This practice is not only clearly visible in the database. The INNO-Appraisal data shows that this practice leads to particularly high quality of evaluations; specifically, the application of open tender procedures is linked to high quality scores.

Third, evaluation reports are very often publicly available and there is particular interest in the evaluation community. The foundation of the Society of Evaluation and several attempts of standardization have intensified scholarly debates. In fact, there is some standardization of approaches visible, but more importantly, this convergent development takes place at a high quality level and includes the openness of evaluators (and commissioners of evaluations) towards new methods.

Finally, we have evidence from the data as well as from expert interviews that learning is a purpose of the commission of an evaluation. There are many formative evaluations, methods such as focus groups or workshops are often employed, and the results of an evaluation are intensively discussed within government. Generally, it seems that learning does take place. However, although we find a high number of accompanying (and interim) evaluations in Germany, it seems that learning applies in fewer cases to the evaluated measures themselves, but takes place on a more general level, namely, the overall policy learning for future policy making and programme design. One of the reasons for this is that the aspect of “policy/strategy development” is an integral part of formative evaluations in Germany.

3.3. UK

It is a widely accepted belief, supported by documented evidence, that the UK has a strong culture of evaluation in RTDI policy making. This case study examines the broader context within which the processes of review, assessment, appraisal, monitoring and evaluation are employed within the UK system of innovation policy governance, a system which, due to the broad definition of innovation held in the UK, encompasses a number of policy domains and actors.

In particular, a number of relevant features of the UK innovation policy governance system are considered, including:

- The use of strategic review processes (and a framework for performance monitoring)
- The presence of multiple actors and stakeholders
- Multi-level governance
- The evolutionary shift from direct support to framework support.
The study then looks at the underlying factors and developments that have shaped the evolution of the current system of evaluation practice in innovation policy governance. These are: a) the development of systematic approach to evaluation in the 1970s and 1980s; b) the accumulation of evaluation expertise through limited meta-evaluation that has led to an innovation culture in government which recognises the value of a practical business oriented approach to policy; c) the growing consensus around the neoclassical model of the economy and society; and d) the extension of evaluation activities throughout government as the devolution of policy and programme and project design and their evaluation has been pushed downwards and outwards from Whitehall to the regions.

Current evaluation practices and tools are then reviewed, in the context of recent structural changes in the machinery of governance in the UK, with a focus on those employed by the Department for Trade and Industry (DTI) and its more recent incarnations, the Department for Innovation, Universities and Skills (DIUS) and now the Department for Business, Innovation and Skills (BIS). The overarching influence of HM Treasury across all policy domains (and the imperative of demonstrating ‘value for money’ from policy intervention) is exemplified by the guiding principles set out in its ‘Green Book’, whilst the promotion of a systematic approach to the policy cycle and to performance measurement (including the use of appraisal, monitoring and evaluation) is underlined by the use of tools such as business cases, programme plans, balanced scorecards and the ROAME-F tool. Evidence is also provided for the cascading down of this guidance to the regional level of governance.

There is also support for the fact that policy interest in the UK extends beyond the mundane and routine application of evaluation as a formalised requirement and into the more exploratory and learning-oriented application of evaluation as an evolving policy tool which is adaptable to a variety of new and changing contexts. This is evinced by the ‘Magenta Book’, which provides guidance on social research methods for policy evaluation and endeavours to develop a greater understanding of the use and applicability of various approaches to evaluation, from the broad to the specific level.

Overall, it is clear that there is an extensive literature and a range of embedded practices relating to appraisal and evaluation in the UK policy system, all of which reinforces the view that the country possesses a well developed evaluation culture.

The study ends with a more detailed examination, in the UK context, of a number of issues which the INNO Appraisal survey of evaluation reports sought to investigate. These were:

- The rationale and purpose for an evaluation: primarily this is aimed at ensuring value for money, coupled with policy learning considerations, which can include identifying unanticipated outcomes and spill-over effects.
- The sourcing and selection of evaluators: all evaluators are external, ensuring independence and evaluation competence, with open tendering a preferred option. Evaluators must meet stringent criteria.
- The use of terms of reference and opportunities for innovative evaluation approaches: Terms of reference are set according to established principles; exploratory approaches are encouraged, provided the principal requirements for the evaluation are met.
• The timing of evaluations: depends on context – the rolling nature of UK programmes tends to favour interim evaluation. Monitoring and appraisal are also standard practices.
• The conditionality of evaluations: evaluation is a pre-condition of HM treasury funding for interventions above a certain funding level.
• The use of dedicated budgets for evaluation: Evaluations are always foreseen and budgeted for.
• Planning of evaluations: All programme formulation includes appraisal, monitoring and evaluation as anticipated elements.
• Topics, data collection methods and data analysis methods: these are all highly dependent upon the context and purpose of the innovation support measure under evaluation. The Magenta Book offers guidance on the appropriate methodologies for use.
• Programme impacts: Evaluations tend to look for both anticipated and unanticipated impacts. Again, the Magenta Book provides guidance on programme impact and how it may be measured.
• Sponsors, audiences and the availability of results: Programme managers form the immediate audience although HM Treasury is the ultimate audience and sponsor. Evaluation in BIS is also under scrutiny from a high level steering group. As a rule, all evaluation reports are made publicly available, except in certain cases where confidentiality concerns arise.
• The production and uptake of recommendations: Recommendations, provided they are realistic and economically feasible are generally acted upon. Similarly, they will be published provided confidentiality concerns do not arise.
• Quality and utility: Quality is defined as being fit for purpose, meeting the Terms of Reference and delivering within budget. Quality is an asymptotic function: there is a minimum level of quality that must be achieved for the delivery of the evaluation's objectives. An evaluation is deemed to be useful if the evaluation delivers the Terms of Reference in a consistent manner and if it provides actionable recommendations and delivers value for money.

In conclusion, it is clear that the UK does indeed possess an extensive and historically well-developed culture of evaluation which though formalised and set firmly in a framework geared towards the assessment of performance measurement, policy relevance and value for money, is nonetheless adaptable, context sensitive and reflexive and, moreover, practised by a policy community that appreciates it as a key tool for policy learning and improvement.

3.4. The case of the Mediterranean Countries (Cyprus, Greece, Italy, Malta, Portugal and Spain)

The aim of this case study is to examine the present situation in the six Mediterranean countries (Cyprus, Greece, Italy, Malta, Portugal and Spain) with regards to the ways evaluations are carried out. It is mainly based on the results of the specific questionnaire survey carried out under the INNO-Appraisal study and more specifically focuses on the evaluation topics covered, the identified data analysis and collection methods, as well as the level of quality and usefulness of the evaluations. These findings are then compared to the overall results of the INNO-Appraisal study in order to examine possible identified inconsistencies and differences.

Given that the evaluations in the countries under the focus of this study are mainly carried out according to Structural Funds requirements, the results are similar to those of the Structural Funds
Executive Summary

(SF) type evaluations examined in the in-depth case study on SF. However, the evidence base is different; all SF type evaluations in the SF case study compared with non SF type evaluations vs. the six countries’ results compared with the total results of the INNO-Appraisal survey.

The initial research hypotheses were as follows:

- Specific evaluation topics are covered in the countries examined vs. the overall results;
- Specific data analysis and collection methods are followed in these countries;
- Specific audiences are addressed;
- Specific quality characteristics are covered;
- Specific issues of usefulness, dissemination and consequences are addressed.

The specific case study mainly draws upon the results of the specific questionnaire template survey carried out under the INNO-Appraisal in comparison with the overall results of the project in order to discover differences and draw substantial conclusions, as well as test whether the aforementioned hypotheses made are indeed the case in the Mediterranean group of countries.

The survey has indicated a small number of differences, but mainly across the different evaluation types, rather than across the Mediterranean countries and the overall population. This suggests that what really makes the difference is SF regulations in terms of how the evaluation types are conducted, but do not seem to suggest anything different from what is usually dictated by international practice, something which is also reflected in the overall results. In terms of quality characteristics, all of them are less satisfactory in the case of the Mediterranean countries in comparison with the overall results. Yet, when examining the results in the six countries in isolation, it is interesting to note that almost all quality characteristics score between 3 and 4 on a 1-5 point scale in terms of satisfaction. This fact can be considered a relatively positive impact of SF regulations given the lack of evaluation tradition in these countries. However, despite the relatively good quality of these evaluations, their results are rarely discussed with government cycles or relevant stakeholders, which is another striking difference with the overall results.

4. Conclusions and Ways forward

This study has, for the first time, provided the policy community and the evaluation community in Europe with a statistical account and analysis of evaluation practice in Europe. Evaluation practice in Europe is highly diverse: it differs between countries and it shows an enormous range in terms of methodological approaches, coverage of topics, quality and usefulness. Different institutional settings and policy traditions in countries influence evaluation practice – and vice-versa, as especially the Austrian case has shown. Evaluation has spread across Europe as the structural fund provisions have pushed countries towards evaluation – though with mixed results to date. The analysis presented in this report constitutes an important step forward in our understanding of evaluation. One key consequence, or so the authors of the study hope, is that the results will allow both policy makers and evaluators to reflect on their own practice, on their approach to evaluation and, ultimately, on the use of evaluation.

While readers may draw their own conclusions as to the lessons to be learned from the analyses presented in this report, and while each of the chapters delivers specific insights from which lessons
can be drawn, there are a set of key observations that should support further improvements in evaluation practice across Europe. Once a rarity, evaluations are becoming increasingly commonplace, yet the analysis has shown that this does not automatically lead to good quality evaluations and productive learning as a consequence of evaluations. Greater care needs to be taken along the whole policy cycle to ensure that evaluations are correctly designed, implemented and utilised, with close interaction at all stages between those commissioning and those performing the evaluations. Policy makers need to be ‘intelligent costumers’, they need to have the absorptive capacity to understand what evaluations can deliver and what they cannot deliver. Evaluators, in turn, must ensure quality throughout the process, especially, though not exclusively, in the application of methods and the development of a thorough understanding of the wider policy and political context in which measures are situated.

Further, conditions and practices concerning the discussion of evaluations within government and beyond must be improved. More thought needs to be given at the planning stage to this phase of the process and to the channels of communication that can be exploited, but evaluators themselves also have to bear in mind that the likelihood and quality of subsequent discussions are highly dependent upon the perceived quality of their reports and the clarity with which methodologies are described and results presented. All this then leads to a more fruitful discussion within and across government and better-informed decisions. In future, however, there will be a need for even greater conceptual clarity given the increasing complexity and sophistication of both innovation policy and the evaluation tools needed to assess the impacts of these developments. The case study of behavioural additionality demonstrated how complex it is to turn one important idea into an operational concept that is both theoretically sound and offers added value to policy makers.

Other operational improvements are also needed. These include the more tailored and conscious design and use of monitoring systems, with evaluations building on the data they produce and monitoring becoming an integral part of the learning process. Evaluation, moreover, should be perceived as a mobilising tool for innovation policy at large, a function highly underused.

Finally, a dilemma confronting evaluation has to be noted. In order to provide the new methods and concepts needed to better inform policy, evaluation itself has to be innovative. Yet the commissioners of evaluations are often very conservative, specifying conventional methodological approaches in their terms of reference despite known limitations and shying away from more experimental approaches. Opportunities to push the boundaries of evaluation theory and practice are thus often constrained.

Allowing for more experimentation, however, will become more important in the future. Evaluation practice in Europe will have to follow the principle of ‘form follows function’ much more closely. The evaluation of innovation policy will have to adapt to new trends in innovation policy and the demands being placed upon it. The analyses in this report have shown a considerable degree of uniformity of evaluation designs across policy measures. Evaluation practice, to a large degree, is an exercise in ‘copy and paste’ into new application areas. However, policy measures are likely to differ even more in the future, and evaluation will have to adapt. To highlight one key example, one major trend is the increasing importance of demand-driven innovation policy and diffusion-oriented measures. For these, evaluation practice is almost non-existent. This has a set of implications. Evaluation will have to tackle systematically and with methodological rigour a broader range of
impacts – the focus on technological and economic impacts is increasingly too limited. Our understanding of how demand-side drivers and policies can interact with and influence supply-side developments also needs to improve radically before adequate evaluation approaches can be developed, and this understanding has to be shared by policy makers and evaluators alike.

A second example concerns the vastly increased emphasis the structural funds place on innovation, where there is a clear need for new innovation concepts in extremely challenging environments. Without the development of intelligent and appropriate evaluation concepts and practices along the policy cycle, there is the danger that new application areas and innovation policy instruments might be supported by evaluation practices that are transferred without any consideration for contextual differences or – even worse – driven by ideological preconceptions. Hopefully, however, the lessons from INNO-Appraisal, the discourse we hope to support and the learning tool we provide can be of some assistance when designing and implementing improved and tailored evaluation approaches that will be needed in the future.

A final – and major – recommendation as to how the results of the study should be used relates to the repository that the study has designed and set up. This repository – in conjunction with the overall statistical data delivered in the study – is a comprehensive authoritative source that documents and codifies practices. The number of policy makers concerned with innovation policy and the number of analysts concerned with its assessment and improvement is constantly growing. Certainly there are guidebooks and manuals that describe evaluation concepts, methodologies and analytical techniques, and there is now an appreciable academic literature on evaluation, but the most numerous and useful sources of information – namely evaluation reports themselves – have to date been firmly embedded (some would say buried) in the relatively inaccessible ‘grey literature’. INNO-Appraisal codifies much of the tacit knowledge that currently exists about evaluation practices and acts as a repository for this knowledge. It thus constitutes a source of learning for newcomers, a reference point for experienced practitioners and one way of helping to overcome problems associated with porous institutional memories. The INNO-Appraisal team strongly recommends that the EU Commission further invests in keeping the repository up-to-date, thus ensuring the survival of an institutionalised learning tool for evaluation and innovation policy in Europe. Moreover, INNO-Appraisal should be seen as a starting point for greater self-reflection by the evaluation community, with many more in-depth studies needed on evaluation practice and its contextualisation.

---

3 First discussions between the INNO-Appraisal team and officials from DG Regio were held on February 4 2010 concerning the transfer and further development of concepts for structural fund evaluations in the area of innovation policy.