

Pilot test of new evaluation methodology of research organisations

Final Report

Vlastimil Růžička, Vladimír Majer,
Tomáš Kopřiva, Petr Vorel,
Hana Bartková,
Andrea Weinbergerová

In 2015, the pilot test of a new evaluation methodology of research, development and innovation, in which twelve research organisations active predominantly in two fields (Chemistry and History) participated, was carried out. Three main and nine subject panels, in which thirty-five foreign and six local experts were present, prepared evaluation reports of thirty-one registered, field-specific research units. The feedback of the panellists and the institutions evaluated, which are useful for the preparation of a nationwide evaluation of research organisations in the Czech Republic, are the key output of the pilot test.

Pilot test of new evaluation methodology of research organisations

Final Report

October 2015

This document has been prepared as a part of the Individual National Project for the area of Tertiary Education, Research and Development „Effective System of Research Financing, Development and Innovation, CZ.1.07/4.1.00/33.0003“. The project was realised by the Ministry of Education, Youth and Sports and financed by the Operational Programme Education for Competitiveness of the European Social Fund and state budget of the Czech Republic.

Summary

This document summarises the results of the pilot test of the methodology of evaluation of research activities (hereinafter the “pilot test”) based on the conclusions and recommendations of the project „An Effective System of Evaluation and Funding of Research, Development and Innovation”. The pilot test was carried out by the “Key Activity 4” team (hereinafter “KA4 team”) of the Individual National Project during 2015. Its objective was to assess all the processes of the proposed evaluation methodology of research activities as suggested in the proposal prepared by the Technopolis Group and its subcontractors, Technology Centre ASCR, Nordic Institute for Studies in Innovation, Research and Education NIFU, and InfoScience Prague (hereinafter “TP”) described in a study “Methodology of evaluation of research and development and funding principles” (hereinafter “Methodology”) and adjusted, taking into consideration the comments from the public consultations, outcomes of the small pilot evaluation and personal experiences of the KA4 team members with comparable evaluations.

The pilot test was realised in full, though the time available and technical conditions to fulfil the task were not optimal. When the pilot test was launched in early 2015, the realisation KA4 team had not had the final reports and recommendations of the previous small pilot evaluation, held on a broader field of science spectrum in the previous year of 2014 and fully realised by the TP; the outcomes of the small pilot evaluation were not published until May 2015. Some methodological approaches of the initial proposal were not feasible due to the time restrictions. After consultation with the teams of the other key activities (KA2 particularly) and the TP, the realization KA4 team therefore proposed certain changes to the approaches suggested.

At the time of launching the pilot test, KA4 team did not have any kind of functional information system at hand. As a provisional solution, a platform, based on agreement with the IT experts, was provided by Czech Technical University of Prague. A simple information support system was provided, which has proved to be functional and absolutely sufficient for the needs of the pilot test. However, for the purposes of future nationwide realisation of the Methodology and before starting the evaluation process, it is necessary to ensure the development and operation of a substantially larger, fully-fledged and functionally validated information system.

A crucial point in the pilot test was identifying the testing sample. For the purposes of the pilot test, the fields of “Chemistry” and “History” were chosen as the main tested Fields of Science. Choosing these two Fields of Science enabled testing the extent to which the new Methodology is suitable for completely different Disciplinary Areas. Only a small group of Research Organisations, as homogenous as possible in their fields to keep the numbers of evaluation panels necessary to a minimum, was asked to voluntarily join the pilot test.

The number of the chosen “Research Organisations RO” that have been addressed and agreed to participate in the pilot test was twelve. The type “Scientific Research Organisations ScRO” was represented by seven faculties of four public universities from Brno, České Budějovice, Pardubice and Prague, and four institutes at the Academy of Sciences of the Czech Republic (AS CR), located in Prague. The type “Industry & Business Services Research Organisations, IBRO” was represented by two private, “Research and Technology Organisations, RTO” located in industrial zones in Ústí nad Labem and Pardubice. The organisation type “National Resources, NatRes” was represented by the National Archive and National Technical Museum, both situated in Prague. Choosing an absolute majority of research organisations based in Prague was due to primarily organisational reasons. Exceptional personnel costs that were carried by the institutions due to their participation in the pilot test were, as a part of the project, contractually covered.

The chosen research organisations acted as “Evaluated Units, EvU”, with the exception of universities, where the Evaluated Unit was, as a priori defined, the faculty. Based on the instructions given, a total of 15 Evaluated Units carried out an internal division into field-specific “Research Units, RU”, of which there were 31. When establishing the Research Units, the internal organisational structure of the Evaluated Units was respected. In several cases, the Research Organisation was identical to the Research Unit. The Research Units were assessed within three Disciplinary Areas. For each Disciplinary Area, three-member main evaluation panels were set up (the Chair - a foreigner, 2 members nominated from within local scientists, and representatives of users of R&D). In the Fields of Science “Chemistry” and “History”, which were the main subject of the pilot test, the evaluation panels consisted of 5 or 6 members and these can be considered, given the proposed Methodology by the TP, to be complete. In other Fields of Science, where the number of evaluated Research Units was between one and four, subject evaluation panels with a reduced number of members were set up (2 to 4 members). The evaluation system had the following structure: Natural Sciences (subject evaluation panel “Chemistry” + 1 reduced subject evaluation panel), Engineering and Technology (4 reduced subject evaluation panels) and Humanities (subject evaluation panel “History” + 2 reduced subject evaluation panels).

Staffing the main and subject evaluation panels that took place during the pilot test was unsystematic, largely due to time constraints, and was based on proposals made in the recommendations of the Technopolis Group, (KA4 team was provided with a list of potential experts), as well as the recommendations of the KA4 team members, and, in some cases, of the institutions evaluated themselves. The aim was primarily to verify the functionality of the system, despite a certain degree of improvisation present when making the important decision of choosing the experts acting as evaluation panels members. The KA4 team worked under disproportionate time pressure, which was manifested by a substantial percentage of addressed foreign experts denying their participation in the pilot test (mostly due to time and capacity constraints). Finally, thirty-five foreign and six local experts joined the evaluation panels and took part in the pilot test. The foreign experts were primarily from “old” EU member states, a smaller part was from South-East or Eastern Europe. Certain complications arose when determining the compensation sums and their taxations and also during reimbursing travel costs. These issues should be paid close attention to in the future and they should be resolved in accordance with standards common in developed European countries.

The evaluation of every Research Unit by the subject panel is based on the so called self-assessment report. As a part of the pilot test, this document was a form including 63 different items. This report, along with the additional materials, became the main source of information for grading every Research Unit on a 5-grade scale (A to E) in five different criteria: *I Research environment and potential; II Membership of the global and national research community; III Scientific research excellence; IV Overall research performance; V Societal relevance.*

In the self-assessment report, the individual research units also performed their own selection of a limited number of outputs of research activities that they themselves considered excellent. For these outputs (this concerned mainly publication outputs of basic research, but also outputs of applied research such as patents, technology, etc.), external reviews were available to the subject evaluation panel members. Due to time constraints, the independent referees of excellent outputs were also chosen unsystematically based on direct proposals of the KA4 team members and so that every excellent output would ideally have two external reviews with a grade on a 5-point scale. This condition, however, was not achieved and members of the panels were asked to participate in the reviews partially during the meetings held in Prague. The attained grades were the main guideline for evaluation in criterion III *Scientific research excellence.*

For every Research Unit, the evaluation panels also had a special bibliometric report prepared by the Key Activity 1 team based on data, obtained from the RIV¹ database and with the subsequent usage of tools that are provided by the *Web of Science*. This report also served as a guideline for the evaluation of criterion IV *Overall research performance*.

In the case of both of the documents, a substantial difference between the fields of exact sciences (Natural Sciences and Engineering and Technology) and Humanities appeared. In fields of exact sciences, both the selection of excellent outputs and the bibliometric report reflected mostly information generated by long-term, field-specific databases and evaluation systems (*Web of Science*, *Scopus*). In Humanities, international systems of this type are practically useless, as they usually do not record and evaluate the main types of outputs, i.e. primarily monographs, moreover usually published in different languages than English. Therefore, “*peer-review*” was much more significant in Humanities.

The self-assessment reports were completed by the evaluated organisations during April, 2015; the bibliometric reports and external reviews of excellent outputs were delivered in May and June. In May 2015, calibration meetings were held (by the main panel and chairs of the subject evaluation panels in all cases). In June 2015, the members of the evaluation subject panels concentrated on the so called remote evaluation, i.e. processing the self-assessment report and bibliometric report and preparing individual evaluations of every Research Unit. The most intensive part of the evaluation process took part at the end of June / beginning of July 2015, when the members of all the evaluation panels personally took part in the main meetings held in Prague, which were approximately one week long in every Disciplinary Area. Information gathered from the self-assessment report, bibliometric report and from reviews of excellent outputs were also combined with on-site visits of the subject evaluation panels at chosen evaluated workplaces. Due to the time and logistic demands, the on-site visit took place approximately at one third of the Research Units. This process enabled verifying how representative the evaluation of Research Units based on solely three types of documentation (self-assessment report, bibliometric report and reviews of excellent outputs), which are key to the Methodology as proposed by the TP, was.

The output of the pilot test was the evaluation of all Research Units in the above mentioned criteria, using a grade and a narrative. Furthermore, at the level of Evaluated Units, the chairs of the main and subject panels collaborated in preparing synthetic evaluation reports in the form of a SWOT analysis. These outputs are not meant to be published and served only to verify the functionality of the whole system. The follow up feedback, which was obtained by the KA4 team members from the evaluators on the one side (both the members of the expert evaluation panels and the referees of excellent outputs), and from the Evaluated Units on the other side, had more value for this purpose.

When evaluating the Methodology as a whole, the opinions of the members of the evaluation panels differed quite significantly from the views of the representatives of the research organisations.

The foreign experts of the evaluation panels have, by a majority, recommended the proposed Methodology as a right move in the direction towards implementing evaluation standards that are already commonplace in many European countries. At the same time, they pointed out that a system based on mechanical rating of outputs with points is rare and does not fully reflect the scientific level of quality of a Research Unit.

¹ RIV: Research Information Register, a central information system for outputs in research and development in the Czech Republic.

The representatives of research organisations evaluated the new Methodology less unanimously, mainly due to its complexity and challenges, however despite certain objections; they evaluated their experience to be mostly helpful. Although some of the Evaluated Units admitted in the follow-up feedback that they had not given the pilot test as much attention as it would have needed if its results had specific, particularly financial, outcomes. Basically, the general usefulness of evaluation of this kind and using its results for research organisations themselves and its management was not doubted. High costs in terms of time and money, combined with the application of the Methodology proposed, as well as a substantial threat of subjectivity within the evaluation system, were assessed as threatening to the financial stability of research and development by the research organisations. This opinion was accentuated even more in the case of Natural Sciences and Engineering and Technology, which have a fully developed and functioning system of international evaluation of quality of publication outputs. As far as Social Science and Humanities are concerned, the new Methodology is considered definitely useful, though it is emphasised that it does not take into account the specifics of various fields enough. The views also vary depending on the type of research organisation. Whereas the university faculties evaluated were mostly critical to the new Methodology, positive opinions prevailed in other organisations. This was particularly significant in the case of institutes of AS CR, which have already had many years of experience with a “peer-review” form of evaluation.

The pilot evaluation has, definitely, fulfilled its objective and proved that the recommendations of the TP, obtained as a part of the IPN Project “An Effective System of Evaluation and Funding of Research, Development and Innovation” are useful as a basis for future evaluations, however, with certain necessary modifications (see Implementation recommendations of the project team IPN Methodology, where a number of conclusions taken from the pilot test are also used). After certain changes and utilising the experience with comparable systems of evaluation, such as “Methodology 2013+”, as well as the currently undergoing evaluation at AS CR, the IPN Methodology team recommends implementing the new system of evaluation as soon as possible, in the first trial phase in only a chosen segment of research organisations, i.e. public universities, with a certain though not crucial impact on the level of their institutional funding. This interim period seems to be necessary, mainly due to the complexity, as well as the cost and time needed for the implementation. After successful completion of this phase, a nationwide system of evaluation and funding of all research organisations in the Czech Republic can be established. The KA4 team is, however, aware of the fact that practical implementation of the new Methodology nationwide is, above all, primarily a political decision.

List of content

1. Introduction	9
1.1 The objectives of the pilot test	9
1.2 Research Organisations, Evaluated Units and Research Units in the pilot test	10
1.2.1 Recommendations for Registering a Research Unit	12
2. On-line Support System	13
2.1 On-line information support	13
2.1.1 Recommendations for the Information System	14
3. Self-assessment Report	16
3.1 Structure and Content of Self-assessment Report	16
3.1.1 Recommendations for the Structure and Content of a Self-assessment Report	17
3.2 Filling Out the Self-assessment Report by Research Units	18
3.2.1 Recommendations for Preparation of the Self-assessment Report	19
4. Setting up the Panels	20
4.1.1 Recommendations for the selection of expert panel members and their roles	23
5. Calibration Exercises	24
5.1.1 Recommendations for Calibration Exercises	24
6. Bibliometric report	25
6.1 Differences in comparison to the TP proposal	25
6.2 Problems occurring during the compilation of the bibliometric report	26
6.2.1 Recommendations for the bibliometric report	26
7. Evaluating excellent outputs	27
7.1 Background	27
7.2 Evaluating excellent outputs and organisational matters	28
7.3 Evaluation results	31
7.3.1 Recommendations for Evaluating Excellent Outputs	33
8. The work of the panels	34
8.1 Remote evaluation	34
8.1.1 Recommendations for remote evaluation	34
8.2 Personal panel meetings in Prague	35
8.2.1 Logistics	35
8.2.2 On-site visits	35
8.2.3 Recommendations for on-site visits	36
8.2.4 Recommendations for the organisation of on-site visits	37
8.2.5 Preparation of evaluation reports for the Research Units and Evaluated Units	37
8.2.6 Field of science and disciplinary area evaluation reports	40
9. Compensation	41
9.1 Panel members	41
9.2 Referees	41
9.3 Research/Evaluated Unit	42
9.3.1 Recommendations for compensation	42
10. Feedback on methodology	44
10.1 Panel members	44
10.1.1 General questions	44
10.1.2 Specific questions	46
10.2 Research and Evaluated Units (RU/EvU)	50
10.3 Referees	55
11. Conclusion	56
12. List of background documents	58
13. List of terms and abbreviations	59

List of Figures and Tables

Figure 1	Pilot Test Schedule.....	10
Figure 2	Qualitative grades given to excellent outputs in disciplinary areas	32
Figure 3	The number of qualitative grades based on criterion I given in the pilot test	38
Figure 4	The number of qualitative grades based on criterion II given in the pilot test	38
Figure 5	The number of qualitative grades based on criterion III given in the pilot test	38
Figure 6	The number of qualitative grades based on criterion IV given in the pilot test.....	38
Figure 7	The number of qualitative grades based on criterion V given in the pilot test.....	39
Table 1	List of Research Organisations/Evaluated Units Participating in the pilot test.....	11
Table 2	Composition of main and subject panels for the pilot test of evaluation methodology of R&D&I	21
Table 3	Panel members by their country of residence	22
Table 4	Total number of approached panel members versus accepted participation	22
Table 5	Example of integrated report on excellent outputs by one panel	30
Table 6	Example of integrated evaluation report on excellent outputs of one research unit RU.....	30
Table 7	Overview of registered excellent outputs divided by disciplinary area and results of evaluation by independent referees	31
Table 8	On-site visits	36
Table 9	Daily remuneration of experts	41
Table 10	Direct and indirect costs of the pilot evaluation.....	42

1. Introduction

This report summarises the results of the pilot test of the research, development and innovation evaluation methodology (hereinafter “Methodology”), which is based on the conclusions and recommendations of the Individual National Project “An Effective System of Evaluation and Funding of Research, Development and Innovation”, realised by the Ministry of Education, Youth and Sports. The report is divided into eleven chapters, of which eight are devoted to describing the individual, successive processes of the Methodology and its pilot test. The before last chapter summarises the feedback of the members of the evaluation panels and Research and Evaluated Units on the Methodology. In the last chapter, general conclusions and recommendations, useful for adjusting and setting the future methodology of research, development and innovation, are stated in concise form. These conclusions and recommendations also reflect the results of discussions of the whole IPN Methodology project team during its external meeting, held on October 1-2, 2015. Specific outcomes and recommendations relating to particular characteristics of the Methodology and the respective processes are listed at the end of every corresponding section or chapter.

1.1 THE OBJECTIVES OF THE PILOT TEST

The aim of the pilot test was to verify all the processes of the proposed Methodology based on the proposal prepared by the Technopolis Group and its subcontractors, Technology Centre AS CR, Nordic Institute for Studies in Innovation, Research and Education NIFU, and InfoScience Prague (hereinafter “TP”). Unlike the small pilot evaluation carried out by the TP at the turn of 2014 and 2015, which was realised using a sample of 17 Research Units chosen based on the prevailing subjects of their sub-field of science research activities and which utilised simplified structures and processes of evaluation, it was our aim to verify all processes as closely and consistently as possible to a situation of a nationwide application of the Methodology.

Roughly 7 months were available for the realisation of the pilot test, while the results of the small pilot evaluation were not published until May 2015, several months after launching the pilot test. Mainly due to a limited scope of time for the realisation, it was decided that only a small number of Research Organisations would be asked to participate in the pilot test, even at the cost of a relatively narrow Field of Science focus. Amongst this small number of Research Organisations, it was attempted to include organisations that would represent, if possible, all four typologies as defined by the TP, i.e. organisations with different missions. The Research Organisations chosen cover institutions of three typologies: firstly, university faculties and institutes of AS CR (“*Scientific Research Organisations, ScRO*”), secondly, private, research and technology organisations (“*Research and Technology Organisations, RTO*”), and thirdly, national resources (“*National Resources, NatRes*”). A common denominator of all these institutions is their sole or partial focus on research in one out of two Fields of Science, Chemistry or History. The choice of these two Fields of Science enabled verifying the extent to which the new evaluation Methodology is suitable for its given purpose both in Natural Sciences and Technology and Engineering on the one hand, and Social Sciences and Humanities on the other hand. The decision to choose Chemistry was determined by mainly two factors. First, the existence of University of Chemistry and Technology Prague, a public university divided into several faculties and therefore providing the option of realising evaluation for one whole Research Organisation in the first typology group, moreover composed of several faculties, i.e. Evaluated Units. Secondly, by including the Chemical Technology faculties of Czech public universities in the pilot test, Biology and some other Fields of Science from the Disciplinary Area of Engineering and Technology were necessarily included.

Schedule of the pilot test is shown in Figure 1.

Figure 1 Pilot Test Schedule

	XII 2014	I 2015	II 2015	III 2015	IV 2015	V 2015	VI 2015	VII 2015	VIII 2015	IX 2015	X 2015
invitation of EvUs to the pilot test											
registering RUs to the pilot test											
elaboration of self-assessment report											
selection and contracting expert panel members											
selection and contracting referees											
evaluation of excellent outputs											
calibration meeting of panel members						14., 20., 21.					
remote evaluation of RUs											
meetings of evaluation panels in Prague							29. 6. – 9.7.				
RU evaluation reports, finalisation & approval											
EvU evaluation reports, finalisation & approval											
elaboration of feedback to Methodology by panellists											
elaboration of feedback to Methodology by RUs/RvUs											

1.2 RESEARCH ORGANISATIONS, EVALUATED UNITS AND RESEARCH UNITS IN THE PILOT TEST

At the end of 2014, the management of several Research Organisations was addressed to voluntarily participate in a pilot test. A total of 12 Research Organisations, i.e. 15 Evaluated Units (full list given in Table 1) took part in the pilot test. The originally registered Faculty of Chemical Technology of the University of Pardubice resigned from the pilot test in February 2015. During the registration process, the Evaluated Units registered a total of 31 Research Units in the pilot test.

The “*Scientific Research Organisation*” category was represented by seven faculties of four public universities in Brno, České Budějovice, Pardubice and Prague, and four institutes of the Czech Academy of Sciences (AS CR), located in Prague. The “*Research and Technology Organisations*” category was represented by two private institutions located in industrial zones in Ústí nad Labem and Pardubice. The “*National Resources*” category was represented by the National Archive and National Technical Museum, both from Prague. Choosing an absolute majority of Research Organisations from Prague was due to mainly organisational reasons.

As the main objective of the project was to test the Methodology, the evaluation results were not published and will not have any consequences on the individual institutions having participated in the test. However, the institutions that took part in the test were asked to provide feedback and comments on the evaluation procedure and its results.

As shown in Table 1, the 31 Research Units registered in the pilot test were evaluated by expert panels in nine Fields of Science. Chemistry was represented most commonly (1.4 nomenclature based on the OECD Fields of Science structure in the Frascati Manual²), with 9 Research Units registered in the pilot test, followed by the Field of Science History and Archaeology (6.1 nomenclature), with 6 registered Research Units. None of the Research Organisations registered any inter-disciplinary Research Units (see *Final Report 1: The R&D Evaluation Methodology*, section 4.8.2).

Table 1 List of Research Organisations/Evaluated Units Participating in the Pilot Test

	Name of EvU	Website of EvU	Type of RO	Nr. RUs in EvU	RUs registered in FoS
NATURAL SCIENCES, ENGINEERING AND TECHNOLOGY	University of Chemistry and Technology Prague - Faculty of Chemical Technology	http://fcht.vscht.cz/	ScRO-HEI	2	1.4, 2.5
	University of Chemistry and Technology Prague - Faculty of Environmental Technology	http://ftop.vscht.cz/	ScRO-HEI	2	2.4, 2.7
	University of Chemistry and Technology Prague - Faculty of Food and Biochemical Technology	http://fpbt.vscht.cz/	ScRO-HEI	3	1.4, 1.6, 2.9
	University of Chemistry and Technology Prague - Faculty of Chemical Engineering	http://fchi.vscht.cz/	ScRO-HEI	2	1.4, 2.4
	Brno University of Technology - Faculty of Chemistry	http://www.fch.vutbr.cz/	ScRO-HEI	4	1.4, 1.6, 2.5, 2.7
	J. Heyrovský Institute of Physical Chemistry of the AS CR	http://www.jh-inst.cas.cz/	ScRO-ASCR	1	1.4
	The Institute of Chemical Process Fundamentals of the AS CR	http://www.icpf.cas.cz/	ScRO-ASCR	4	1.4, 2.4, 2.5, 2.7
	Centre for Organic Chemistry Ltd.	http://cocitd.cz/en/	IBRO- RTO	1	1.4
	The Research Institute of Inorganic Chemistry, Inc.	http://www.vuanch.cz/	IBRO-RTO	1	1.4
HUMANITIES	The University of Pardubice - Faculty of Arts and Philosophy	http://www.upce.cz/	ScRO-HEI	3	6.1, 6.2, 6.3
	The University of South Bohemia in České Budějovice - Faculty of Philosophy	http://www.ff.jcu.cz/cs/web/ff/	ScRO-HEI	2	6.1, 6.2
	The Institute of History of the AS CR	http://www.hiu.cas.cz/en/	ScRO-ASCR	1	6.1
	The Institute for Contemporary History of the AS CR	http://www.usd.cas.cz/	ScRO-ASCR	2	6.1, 6.3
	The National Technical Museum	http://www.ntm.cz/	NatRes	1	6.1
	The National Archives	http://www.nacr.cz/	NatRes	2	1.4, 6.1

Explanatory notes:

Type of Research Organisation: ScRO-HEI scientific research organisation, university; ScRO-ASCR scientific research organisation, Academy of Sciences of the Czech Republic; IBRO-RTO industry and business services research organisation, research and technology organisation; NatRes national resources.

Field of Science ("FoS"): 1.4 Chemistry; 1.6 Biology; 2.4 Chemical Engineering; 2.5 Materials Engineering; 2.6 Environmental Engineering; 2.9 Industrial Biotechnology; 6.1 History and Archaeology; 6.2 Languages and Literature; 6.3 Philosophy, Ethics and Religion.

The registration rules for the Research Units corresponded with the TP recommendations, i.e. every Evaluated Unit could register only one Research Unit to be evaluated by the subject panel and every researcher could be included in only one Research Unit (see *Final Report 1: The R&D Evaluation Methodology*, chapter 4). However, the rule of a minimum threshold of at least 50 outputs in the evaluated period was not kept. This enabled the Centre of Organic

² Revised Field of Science and Technology (FOS) Classification in the Frascati Manual, DSTI/EAS/STP/NESTI(2006)19/Final, Feb-2007, OECD

Chemistry Ltd., a recently established private research organisation, to participate in the test, representing the research organisation category IBRO-RTO.

It was recommended the Research Organisations register Research Units that would, if possible, reflect the internal organisational structure of the organisation. This recommendation was respected by the Research Organisations, the Research Units registered corresponded with the institutes (departments) or their aggregates at the public university faculties or sections at the AS CR institutes. In several cases, all the researchers from the Evaluated Units were not included in the Research Units registered, e.g. due to the short existence of the workplace and so far insufficient number of outputs.

1.2.1 Recommendations for Registering a Research Unit

1. Identify a minimal threshold value, e.g. 50 outputs for the registration of a Research Unit RU, with the option of increasing the relevancy of large outputs (e.g. books in Social Sciences and Humanities).
2. Do not cap the size of a Research Unit RU (by a maximum number of researchers or outputs).
3. The Evaluated Unit EvU has the option of registering more than one Research Unit RU in one Field of Science if it can justify doing so (higher number of scientific workers and clear difference in topics).
4. A Research Unit RU should clearly correspond with the organisational structure of the part of the Evaluated Unit that it represents.
5. The Evaluated Unit EvU does not necessarily need to include all its scientific workers in the registered Research Units RU, however, it must inform the panels of the reasons for doing so and of the numbers of scientific workers that were not included.

2. On-line Support System

2.1 ON-LINE INFORMATION SUPPORT

Already prior to the pilot test it was apparent that there will be a great need to acquire vast amounts of data in order to carry out evaluations of Research Organizations, Evaluated Units, and Research Units, and that information and administrative support will also be needed for the actual process of pilot test. The TP was aware of the importance of information and administrative support (see *Final Report 1*, Section 5.2.2), however, the issue was not discussed any further. During the Small Pilot Evaluation, the TP used tools (Excel files) and means of communication (e-mails), which the KA4 team evaluated as less effective for the purposes of the pilot test.

Therefore, the KA4 team presented their own analysis of the issue of information and administrative support of the evaluation process of research organizations in the Czech Republic according to the Methodology. The analysis describes the features of the information system and the requirements for its function. Consultations with KA1 experts confirmed that the IPN Methodology project does not account for sufficient amount of time in order to build a fully-fledged information system that would ensure information and administrative support for the pilot test and that the project budget does not allocate sufficient amounts of funds for an information system that would provide information and administrative support. The existing information systems of internal grant agencies, universities, and the information support system for evaluation of the Academy of Sciences were examined. The information system of the internal grant agency of Czech Technical University was identified as the closest and easiest to use, if it were modified for the purposes of pilot test. Consultations with the creators and administrators of the information system resulted in the conviction that even a modification of the system would be time consuming and the deadlines for pilot test could not be met. Nevertheless, the CTU experts proposed an option to provide the pilot test with information support at least to a limited extent, given the fact that the information support does not encompass all the desired features of a fully-fledged information system. Prior to the onset of the pilot test, a group of IT experts from CTU presented a process analysis and proposed a structure and hierarchy of libraries and directories. The IT experts also secured the operation of the information support during the course of the pilot test.

The information support used modified functions of the Share Point tool. After manual allocation of differentiated access rights, it allowed the sharing of source materials on-line using common standard software to all the participants of the pilot test that needed access (in rare cases, due to the large extent of data, there were problems with the transmission of data from external word processors and spreadsheet editors). In view of the involvement of foreign experts, all texts in information support were in English.

Each Research Unit was appointed a separate directory in the base library “*Documents*”, where the RU stored “*Excellent Outputs*”, and where self-assessment report forms were allocated in hierarchical directories. Filling out forms did not require installing any special software; the default functions of the Web App were sufficient. Since the system that was used was not a fully-fledged information system, the forms lacked control functions, which placed increased demand on the people who entered the data for the Research or Evaluated Unit, as well as on KA4 team members, who carried out the data consistency checks. Setting up the final self-assessment report also required intervention of an IT expert (creating routines combining individual parts and eliminating redundant auxiliary texts). The research organizations, Evaluated Units and Research Units that voluntarily participated in the testing were not required to verify the data with an electronic signature or other form of guarantee of accuracy and integrity of the data.

Another group with separate access rights, the referees of excellent outputs, was provided with a separate folder. The separate folders of the referees included excellent outputs from their Field of Science. The excellent outputs had to be entered into the above mentioned directories manually by a KA4 expert.

A special library was created for members of individual subject panels and each subject panel had its separate folder. The folder contained subfolders of individual Research Units with self-assessment reports, excellent outputs, bibliometric reports and other work materials for evaluation reports. Panel members stored proposals of evaluation reports in directories for individual Fields of Science. The handling of these files was carried out manually. The members of the main panels were granted access rights to all files of the fields belonging to their “*Disciplinary Area*”. The panel library also contained a folder for summary reports at the level of the Evaluated Unit, and a folder with instructional texts, guidelines and general documents prepared by the KA4 team for the use of Research Units, referees and panel members.

The pilot test information support served its purpose and complied with main requirements of information and administrative support – on-line access that does not require any special software, and protection of data confidentiality. The statement above was confirmed by evaluation panel members in their feedback to the Methodology and pilot test. Since the system used was not a fully-fledged information system, the information support lacked some functions, specifically an automated control of entered data, automated settings of access rights, transfer of data and texts into different forms or texts, etc. The self-assessment report forms were not linked to the help function and quotations in the guidelines and instructions texts. The fulfilment of the information support’s purpose was also facilitated by the relatively small scale of the pilot test; only a limited number of research organizations in three disciplinary areas were addressed. In case of a larger number of Research Units, such arrangements of information support would require a far greater number of service personnel and time to transfer and adjust the data.

2.1.1 Recommendations for the Information System

Based on operating the on-line support system, the following recommendations for information and administrative support of the evaluation process have been made:

1. The fully interactive electronic information system (hereinafter “IS”) must be a complete system including all the steps of the evaluation process.
2. The information system must be an original system developed for the purposes of evaluation of R&D&I in the Czech Republic. Modification of one of the existing systems or commercial systems wouldn’t provide the expected quality of information and administrative support. The IS must ensure data transfer from the Czech R&D&I Information System.
3. The information system will be fully autonomous and its users will not be required to use any additional software, the IS will be independent of any browser, it will allow converting, export and import of texts and tables from all common word processors and spreadsheet editors, it will enable sharing and transfer of data and texts amongst different modules.
4. The information system must guarantee data protection against unauthorized use, loss or alteration, and allow storage of data for the necessary control period (approximately 10 years, possibly more).
5. The information system will include a module used for automated generation of access of different user roles.

6. The information system will enable automatic generation and sending of e-mails and contractual documents, and will allow the use of electronic signatures for binding data verification.
7. The information system will ensure monitoring of all access and modification of data (a journal) protected against any adjustment and will include an interactive helpdesk.

We recommend that the information system includes: the supply of licenses incl. maintenance, implementation of IS functionalities, documentation of IS settings for administration (operating system settings, access rights settings, backup, failure recovery, maintenance of system logs, etc.) and user documentation, training of the IS personnel carried by evaluation administrative support workers. Creating an autonomous, robust information system is a necessary condition for the satisfactory conduct of the evaluation process and its monitoring, control and future long-term development.

3. Self-Assessment Report

3.1 STRUCTURE AND CONTENT OF SELF-ASSESSMENT REPORT

In January 2015, the TP provided the KA4 Team with *Guidelines for the evaluated research organizations* for the purposes of the pilot test. The Guidelines were prepared on the basis of experience with the Small Pilot Evaluation. The 27 pages long document was very different in comparison to the previously used guidelines *Small pilot evaluation – Submission guide lines* from October 2014. Both documents focused on a sequence of questions; their number was increased from 44 to 63. Most of the questions concerned the quantitative data that define Research Unit (RU), and were in some cases accompanied by a descriptive text. The questions were set in a new order. After the initial 8 questions that related to RU generally, their succession (Q009 to Q063) corresponded with the five criteria used to evaluate RU. This was significant progress in comparison to the previous unclear guidelines, nevertheless, the KA4 Team did not consider this document entirely suitable to be passed on to the individual RUs as guidelines for the preparation of the Self-Assessment Report. The issue here was its comprehensibility by both the RUs creating the report as well as the panel members evaluating it. Therefore, several methodological or formal modifications were made (after partial consultation with the KA2 Team), but the overall nature of the document was not significantly changed. Hereunder are the main modifications of the final draft of the document that was passed on to the RUs (see **Background document 1**).

1. In the first extensive section regarding the *Research environment and potential* criterion, questions Q009 to Q034 were rephrased, as it was deemed necessary that the information is presented either at the Evaluated Unit (EvU) level, or simultaneously at the level of RUs and EvUs, according to the character of the questions. It is primarily because the field-specific RUs are usually formed ad hoc for evaluation purposes and may not correlate with the organizational structure of the actual EvU. Therefore, some questions, for example regarding institutional funding or infrastructure, cannot be answered only at the level of RU and it is necessary that the evaluating panel members see the clear connection between the RU and EvU. The KA4 team also emphasized the preference of creating RUs in respect to the EvU organizational structure; an organigram of the EvU was required as a part of the Self-Assessment Report.
2. According to the KA4 Team, the original proposal also lacked a short description of the work that was carried out by the RU during the evaluation period and its results, therefore question Q014 *Historical background* in the TP's proposal was replaced by *Description of research activities, objectives and achievements over the evaluated period of the EvU*, with emphasis on the contribution of the given RU.
3. The tabular presentation of data on human resources was homogenized for different types of Research Organisations (universities, institutes of the Academy of Sciences, and others), with an emphasis on differentiating the core scientists from supporting personnel (technicians) and postgraduate students. In order to standardize the results with respect to the actual time spent on research, a coefficient of 0.5 was arbitrarily used to calculate FTE of scientific work of academics from universities, with the assumption that the time was evenly split between scientific and educational work.
4. The tables on postgraduate students registered in the EvU and actually educated in individual RUs were also made more transparent.

5. Contrary to the original proposal, RUs were requested to present a list of selected excellent outputs, and to provide short commentaries explaining why the selected outputs are considered most significant.
6. The required range of data and their number presented in the form of a table were correlated with the size of the RU where appropriate, using FTE of scientific work as a standardizing factor.

Despite the modifications that were carried out under significant time pressure, the KA4 Team was aware that the required content of the Self-Assessment Report and its structuring were not ideal and the questions were not always clearly articulated. The next section describes how the above mentioned concerns manifested during the process of filling-out forms on-line. Also, some panel members, but mainly most of Evaluated Units, had critical comments, see 10.1.2.3. and 10.2.1.1 for their synthetic summary.

Especially the structuring of more than 63 coordinate questions that are not grouped into blocks, which would provide a certain hierarchy, seems to create a rather unfortunate and visually bureaucratic format. The Self-Assessment Report focuses on the summary of the various macroscopic data describing the RU as a whole, while not emphasizing in any way the contribution of individual workers or their groups, thus creating a “black box” impression. Missing list of personnel and publications were among the most common objections. Therefore, it is unclear whether the RU contains excellent scientists or less productive workers. During the calibration, the panels requested a list of all published books of RUs in humanities. There were also multiple requests to divide the report into the opening part of the narrative, which explains the mission, goals and strategies of the RU and briefly describes its activities during the evaluated period, including an outlook of the upcoming period. The second part was to contain all additional quantitative data in a tabular form. The presentation of the postgraduate student’s contribution was problematic, since panel members did not understand the complex system of the Czech Republic and the data, full of mistakes created when the data was entered by the RU, was considered misleading. Also, RUs had reservations regarding the efficiency of the SWOT analysis that was required at the end of the Self-Assessment Report. RUs did not have a unified opinion on whether it should be presented at the level of RU or EvU. The extent to which the presented form of the Self-Assessment Report is suitable to carry out evaluation of the whole EvU is debatable. The key qualitative data in the Self-Assessment Report (number and FTE of scientists, number of postgraduate students, and number of outputs) was transferred into overview summary tables that were included in the Guidelines for Panel Members.

An example of an “empty” Self-Assessment Report is attached (see **Background document 2**).

3.1.1 Recommendations for the Structure and Content of a Self-Assessment Report

1. The Self-Assessment Report will be divided into two sections; section I descriptive part, section II data.
2. The descriptive part will consist of: summary - mission, goals, strategy (1 page), report on activities during the evaluated period (3-4 pages), outlook of the upcoming period (2-3 pages).
3. The data part will consist of blocks divided according to the five criteria; in case of criterion “Research Environment and Potential”, Research Unit, RU will be presented in the context of the Evaluated Unit, EvU.
4. A list of all outputs with links to their authors will be presented as well as a list of core scientists in the given period, distinguished from those who were no longer members in the time of the evaluation, or have become members only recently.

5. Some data (for example the number of outputs) will be standardized to the number of FTE of core scientists. It will be necessary to establish clear rules of calculating the FTE of faculty members at universities (trivial solution – FTE equals one half of physical working time) and for organizations with workers that carry out scientific work only partially.
6. Clear statement of the number of postgraduate students actually working in a research unit.
7. Success rate of postgraduate student's trained in the RU (number of defended theses) and actual average length of study of postgraduate students working in the RU.
8. In the outputs of the Research Unit, include and label works of current or former postgraduate student authors (a specific feature of social sciences and humanities, where the results of doctoral theses are published by the postgraduate student without stating the name of his/her supervisor).

3.2 FILLING OUT THE SELF-ASSESSMENT REPORT BY RESEARCH UNITS

During the preparatory phase of the pilot test (February and March 2015), informative materials were prepared for the participating research units (namely the *Submission Guidelines for the Evaluated Research Organisations*, see **Background document 1**), and after a working meeting that was held for the purpose of becoming acquainted with the objectives of the Pilot Test (February 10, 2015), a working meeting for the responsible and contact persons from Research Units took place (April 14, 2015). Furthermore, the guarantor of KA4 was in permanent contact with representatives of research units and telephone numbers and e-mail addresses of other KA4 expert members were available.

The most common shortcomings included the forms being filled-out incompletely, information inaccuracy and inconsistency of data, failure to meet the deadlines for filling-out and completing the Self-Assessment Report. In the next evaluation, attention must be paid to coordinating when the forms of the Self-Assessment Reports are filled out, particularly in cases where the research organization or evaluated unit (RO or EvU) consists of several Research Units, and to the length of the texts and focus on the meaning of questions. Every so often, the panel members were provided with very short answers that did not sufficiently inform on the assessed feature or activity of the Research Unit for the purposes of evaluation. Excessively long answers with large amounts of unnecessary data and information that did not answer the question were also not an exception (see section 10.1, comments of panel members).

Some of the errors made upon entry of the figures were caused by the information support in use, which lacked control mechanisms for formatting and checking the figures.

The biggest problems that occurred when filling out, but also when being interpreted by panel members, were related to questions on the number and means of involvement of post gradual students in R&D&I activities. In future evaluations, these questions should be precisely worded and structured. It would be useful to follow up on data on postgraduate students that are already available in the register of students.

3.2.1 Recommendations for Preparation of the Self-Assessment Report

1. In terms of the information system, incorporate control mechanisms for filling out individual questions in the form, check the length of texts and the consistency of figures.
2. Make as much as possible use of the data that had already been processed in other information sources, namely use the student register and RIV, and the R&D&I Information System.
3. In determining the evaluation timetable and processing time of the Self-Assessment Report, consider the specifics of the academic environment (for example examination period) and generally known potential sources of time inconsistency (e.g. activity reporting period, or the holiday season, etc.).

4. Setting up the Panels

Under the proposal of the TP, the appointment of panel members and referees is a step-by-step, hierarchical process that begins with the appointment of the main panel chairs, based upon the proposal of the Council for Research, Development and Innovation³.

Since the pilot test had to be implemented in a short time of roughly 7 months, it was decided that the process of selection and appointment of main and subject panel members and referees must be substantially faster than the period required for the proposed hierarchical process. The chairs of the main and subject panels, members of the main and subject panels, and referees were selected by the KA4 Team. The team used several databases of experts who participated in the comparable assessments of research institutions in different countries. The databases included experts registered by the Technopolis Group, members of main and subject panels from the UK RAE2008 and UK REF2014 evaluations, experts from the 2013 Portuguese evaluation that was organized by the Portuguese Foundation for Science and Technology, experts from the evaluation of Royal University in Stockholm, experts from the evaluation of Helsinki University and experts for evaluation of international programmes used by the Ministry of Education, Youth and Sports. Some panel members were selected based on the KA4 Team's request for recommendations of the Evaluated Units. The Czech main panel member positions, representing the Provider or the Ministry, were not appointed. The Deputy Prime Minister Bělobrádek's department of the Office of Government withdrew their initial nomination of the three members, stating that the work description of the main panel members does not correspond with the work description at their main workplace.

The process of addressing and contracting potential experts started in January, 2015 and finished at the beginning of June, 2015. In the last stage of the process, Technopolis Amsterdam's recommendations were used to appoint the remaining missing experts in some subject panels.

Preventing conflicts of interests of panel members and referees of excellent outputs in relation to the evaluated institution is a major part of the assessment. The areas of conflicts of interest and the preventive regulations are defined in the TP⁴ documentation. In terms of the pilot test, we decided to examine the conflict of interest only superficially and we retreated from writing a declaration of absence of conflict of interest. In one case, we have excluded an expert that was proposed by the Evaluated Unit due to her previous employment in the research organization.

Overall, it is clear that the selection of the panel members and referees of excellent outputs was, as far as the pilot test is concerned, improvised, since no practically applicable methodology was available at the time of the onset of the pilot test, even though it should have been established according to the above mentioned basic articles of the evaluation process. A similar process of establishing evaluation committees in "Methodology 2013", which was used in the course of 2014⁵, could be used in the future (because it works across all scientific disciplines).

³ See Final report 1: The R&D Evaluation Methodology, section 5.1.2 *Staffing of the panels*

⁴ See Final report 1: The R&D Evaluation Methodology, section 5.3.1 *Integrity of the panel evaluation process*

⁵ The selection of referees (domestic and foreign) was based on nominations of research organizations. Therefore, only those considered to be renowned experts of the given field (at least at the level of nominating research organization) could become referees. The Committee for Evaluation of Results of Research Organisations and Completed Programs (R&D&I Council) created several working teams for the sub-field groups for the purposes of establishing expert panels. Those working teams were minutely familiar with the research organizations' nominations of expert panel members, and created lists of candidates, so called Primary Proposals, with regard to quality, the field, regional and sectoral representation. The

Table 2 indicates the nominal composition of twelve panels involved in the pilot test, Table 3 lists the panel members according to their country of residence, and Table 4 summarizes how many experts were addressed to be appointed in individual panels. Approximately 70% of the panel members had experience with similar evaluation of research and research organizations.

Table 2 Composition of main and subject panels for the pilot test of evaluation methodology of R&D&I

OECD Name	Surname Name	Affiliation, Town, Country
1. NATURAL SCIENCES main panel	Thulstrup Erik	Roskilde University, Roskilde, Denmark
	Němeček Zdeněk	Charles University, Prague, CR
	Rejholec Václav	Consultant for pharmaceutical industry, Prague, CR
1.4 Chemical Sciences	Hapiot Philippe	CNRS- University of Rennes 1, Rennes , France
	Guillon Daniel	CNRS - University of Strassburg, Strassburg, France
	Haines Michael	Cofree Technology Ltd, Bricklehampton, UK
	Heintz Andreas	University of Rostock, Rostock, Germany
	Kukhar Valery	Nat. Academy of Sciences of Ukraine, Kyiv, Ukraine
	Rizzi Andreas	University of Vienna, Vienna, Austria
1.6 Biological Sciences	Driessen Arnold JM	University of Groningen, Groningen, Netherlands
	Elska Ganna	National Academy of Sciences of Ukraine, Kyiv, Ukraine
	Rodger Alison	University of Warwick, Warwick, UK
2. ENGINEERING and TECHNOLOGY main panel	Seville Jonathan Peter Kyle	University of Surrey, Guildford, Surrey, UK
	Hanika Jiří	Czech Academy of Sciences, Prague, CR
	Souček Ivan	University of Chemistry and Technology, Prague, CR
2.4 Chemical Engineering	Lapicque François	CNRS-ENSIC University of Lorraine, France
	Grievink Johan	University of Technology, Delft, Netherlands
	Ocone Raffaella	Heriot-Watt University, Edinburgh, Scotland, UK
2.5 Materials Engineering	de With Gijsbertus	Eindhoven University, Eindhoven, Netherlands
	Drillon Marc	CNRS - University of Strassburg, Strassburg, France
	Katgerman Laurens	University of Technology, Delft, Netherlands
	Salmi Tapio	Åbo Akademi, Åbo (Turku), Finland
2.7 Environmental Engineering	Rulkens Wilhelmus Henricus	Wageningen University, Wageningen, Netherlands
	Legube Bernard	University of Poitiers, Poitiers, France
	Sánchez Hervás José María	Unit for Energy Valor. of Fuels and Wastes, Madrid, Spain
2.9 Industrial Biotechnology	Voragen Fons	Wageningen UR University, Wageningen, Netherlands
	Jelen Henryk	Poznań University of Life Sciences, Poznań, Poland
6. HUMANITIES main panel	North Michael	University of Greifswald, Greifswald, Germany
	Ledvinka Václav	Prague City Archives, Prague, CR
	Pešek Jiří	Charles University, Prague, CR

proposals were arranged in descending order and the Committee for Evaluation used them to select a corresponding number of nominations for the expert panels.

OECD Name	Surname Name	Affiliation, Town, Country
6.1 History and archaeology	Hadler Frank	University Leipzig, Leipzig, Germany
	Catalano Alessandro	University of Padua, Padua, Italy
	Hengerer Mark	Ludwig-Maximilians-Universität, Munich, Germany
	Mayer Françoise	Paul Valéry University, Montpellier, France
	Müller Leoš	Stockholm University, Stockholm, Sweden
6.2 Languages and literature	Achard-Bayle Guy	University of Lorraine, Nancy, France
	Balogh Andras	Babeş-Bolyai-Universität, Cluj-Napoca, Romania
	Raynaud Savina	University Cattolica del Sacro Cuore, Milan, Italy
6.3 Philosophy, ethics and religion	De Roover Jakob	Ghent University, Ghent, Belgium
	Müller Daniela	University of Nijmegen, Nijmegen, Netherlands
	Thomassen Einar	University of Bergen, Bergen, Norway

Table 3 Panel members by their country of residence

Country of residence	Number	Country of residence	Number
Belgium	1	Norway	1
Czech Republic	6	Poland	1
Denmark	1	Austria	1
Finland	1	Romania	1
France	7	United Kingdom	4
Italy	2	Spain	1
Germany	4	Sweden	1
Netherlands	7	Ukraine	2
TOTAL			41

Table 4 Total number of approached panel members versus accepted participation

Disciplinary Area, Field of Science	Number of approached	Accepted	Accepted in %	Rejected
Main Panel Natural Sciences	5	3	60	2
Main Panel Engineering and Technology	8	3	38	5
Main Panel Humanities	6	3	50	3
History and archaeology	12	5	42	7
Languages and literature	7	3	43	4
Philosophy, ethics and religion	9	3	33	6
Chemical Sciences	25	6	24	19
Biological Sciences	4	3	75	1
Chemical Engineering	7	3	43	4
Materials Engineering	8	4	50	4
Environmental Engineering	14	3	21	11
Industrial Biotechnology	19	2	11	17
TOTAL	124	41	33	83

4.1.1 Recommendations for the selection of expert panel members and their roles

1. Subject panel members will be nominated by the institution responsible for the evaluation.
2. Czech experts will be full members of the subject panel, they are represented in a maximum ratio of one Czech panellist to two foreigners, the panel chair is always a foreigner who makes sure that conflicts of interests amongst Czech panel members are avoided. The panel operates without advisors.
3. Subject panel members must have expertise to evaluate sub-fields („*Sub-field of Science*“) that are the dominant research activity of the participating research unit.
4. The number of panel members will correspond to the number of evaluated research units: 6 panel members for less than 15 research units, 9 panel members for 15 to 25 units, 12 panel members for more than 25 units. Large panels will operate in two or three parallel sections. A panel should not evaluate more than 40 units.
5. Each unit will be assessed by three panel members (typically one Czech and two foreign members) during remote evaluation; the evaluation time will vary from a half to one full day, depending on the size of the unit.
6. The time specifically designated to evaluate one unit during personal panel meetings is typically within 4 hours.
7. The expected number of panels in the evaluation of all research organizations in the Czech Republic ranges from 30 to 35, two or three panels can be appointed in the major fields (e.g. biology or physics), but panels can be, in the case of small fields, merged.
8. Precisely define the roles and responsibilities of members of the main panels.
9. Explain the complex system of postgraduate studies in the Czech Republic to foreign panel members.

5. Calibration Exercises

The description of calibration exercises is mentioned several times in the documentation of TP, namely in reports number 1 and 3 (*Final report 1: The R&D Evaluation Methodology*; *Final report 3: The Small Pilot Evaluation and the Use of the RD&I Information System for Evaluation*), in the Summary Report, and in the Evaluation Handbook (*Background Report 5: Evaluation Handbook*).

Considering the short time for implementation of the pilot test, we decided to limit the calibration to only one meeting, and invite the main panel chairs and members and subject panel chairs or their substitutes. Since there were three main panels for three different disciplinary areas, in total three meetings took place in the first half of May 2015. In the first part, the participants got acquainted with the basic principles of the Methodology and with the pilot test goals. The second part was dedicated to a discussion on calibration settings of individual sub-criteria in criteria of “*Research environment*”, “*Membership in national and global research community*”, “*Overall research performance*” for all four types of research organizations. The terms “*originality, significance, rigour, and reach*” were also discussed. The participants of all three meetings agreed that a field-specific interpretation of those terms does not exist. The subject panel chairs, or their representatives, were instructed to share results of the calibration exercise with the members of their panels.

Minutes of the calibration meetings are provided in a separate document, see **Background document 3**.

The pilot test results showed that it is necessary to pay proper attention to the calibration process, and that it is desirable to develop and clarify the recommendations contained in the TP documents. In our opinion, the weakest feature of the evaluation is a different understanding and interpretation of the grades amongst the subject panels, rather than amongst the panel members within the subject panels themselves, which results in vast differences in evaluation of qualitatively comparable research units (see also feedback of EvUs in chapter 10).

5.1.1 Recommendations for Calibration Exercises

1. Clearly define the number, agenda and the mandatory participants of the calibration meetings.
2. Clearly define the role and tasks of the main panel chair and the chairs of subject panels during calibration exercises and during implementation of the results of these meetings.
3. Clearly define the means to ensure the panel members uniformly interpret the qualitative grades of evaluation according to all five criteria. Uniform interpretation must be ensured amongst subject panels within one disciplinary area as well as between different disciplinary areas (different main panels).
4. Reassess the necessity of field specification of terms as established in the TP proposal:
 - *significance, originality, rigour* in criterion III “*Scientific research excellence*”
 - *reach, significance* in criterion V “*Societal relevance*”.
5. Reassess the necessity of setting the relevancy of the sub-criteria in criteria I “*Research environment*”, II “*Membership in national and global research community*”, and IV “*Overall research performance*”.

6. Bibliometric report⁶

Detailed description of changes, differences in calculations of some indicators, and problems occurring during the compilation of the bibliometric report is given in the **Background document 4**.

6.1 DIFFERENCES IN COMPARISON TO THE TP PROPOSAL

In comparison to the proposal provided by TP, one major change was the addition of several comments providing explanations that should aid in correct interpretations. Further, the years the outputs were from were included into all the graphs and tables (due to the fact that citation indicators were calculated only in 2009-2012).

Further changes were made to the format of the individual tables and graphs.

An example of an anonymous bibliometric report is given in a separate document (see **Background document 5**).

Calculation of some bibliometric indicators differed from the small pilot evaluation. The procedure of calculation was discussed in meetings of the KA1 team and bibliometrical team of the TP. Changes were introduced in indicators A9, C4, and D1.

Apart from bibliometric reports, bibliometric overviews of excellent outputs which were chosen for peer-review evaluation of the referees were also prepared. These overviews contained 5 bibliometric indicators:

- Number of citations – citation numbers;
- Category expected citation – the average number of citations of documents of a similar type, the year they were published in and the field (in the case of a multidisciplinary document, the average value was used);
- Citation percentile – percentile rating based on number of citations in documents of the same kind, year published and field (in multidisciplinary cases, the best, i.e. lowest value is taken into account);
- Journal impact factor – average citation of the journal;
- Journal percentile – percentile rating of a journal based on Journal IF within the same publishing year and field (in multidisciplinary documents, the average value is taken into account).

These indicators and how they are calculated was determined based on consultations with the bibliometric team of the provider's consortium, members of KA4 team and also given the availability of the indicators.

Summary overviews containing data based on five chosen indicators, C1, F4, C4, D1 and D2, were created for panel members' use.

⁶ This chapter was written by Martin Lhoták, Pavel Mika, and Jakub Sczarec, Key activity 1 team members

6.2 PROBLEMS OCCURRING DURING THE COMPILATION OF THE BIBLIOMETRIC REPORT

Out of 3782 researchers (listed as local research outputs authors in the RIV register), 658 remained unclassified in any of the Research Units. In some cases, the reason for the lack of classification was given (member of a different institution, former worker), however, the question remains why they were listed as local authors of the output.

The process of mapping the records in R&D&I information system (RIV) and Web of Science is key when preparing the data. In order to achieve greater efficiency and mainly credibility, it would be useful to use a higher degree of control. Also, given the ethics of evaluation, the participants should have the option to check and confirm the data that their evaluations will be based on.

A closer interconnection of the Czech R&D&I information system RIV with citation databases, as described in the *Background report 9: The RD&I Information System as an information tool for evaluation*, should improve the process of mapping.

Several errors occurred in the calculations during the data processing and filling in the templates of the individual reports. These were not system errors in data or algorithm calculations, but, in most cases, these were caused by human error due to the information load and new ways of processing the data. Most errors were identified when the report was published – together with the corrections, some additional comments and labels of tables and graphs were also included.

For practical reasons, it is necessary to make the process of redistributing information in the tables and graphs more efficient, i.e. change it into text form. Filling in information into the templates manually proved to be time consuming, error prone and made further formatting changes more difficult.

6.2.1 Recommendations for the bibliometric report

1. The bibliometric report will correspond roughly to the proposal of Technopolis Group and its subcontractors, including the changes as proposed in this chapter. An option to simplify the report (recommendation of panel members) should be considered.
2. Czech Republic (CZ) indicators: to be stated for both the field of science in which the Research Unit is registered and the average value throughout fields of science based on the publication profile of the RU.
3. Presentation of key indicators of the bibliometric report also in form of a clear overview table, throughout the RUs.
4. Publications of authors from several RUs should be accounted for multiply.
5. Reflect on evaluation of social sciences and humanities outputs more deeply.
6. Expand the Czech R&D&I information system RIV by book reviews.
7. Resolve the question of taking non-scholarly outputs of applied research into account (already done in pilot test).
8. Bibliometric reports will be generated automatically by the system and available on-line, on a website used for the evaluation.

7. Evaluating excellent outputs

7.1 BACKGROUND

The main input for the evaluation of research units by the subject panels, based on the criterion III *Scientific Research Excellence* in the Methodology, is evaluation of a limited number of selected publications by independent referees from abroad.

Based on the proposal of the TP, each unit presents 1 to 2 percent (and not less than 3 and not more than 20) of its scholarly outputs in English only (*papers in peer reviewed journals, conference proceedings, monographs, books and chapters*) for remote evaluation by two referees. Referees are chosen based on expertise, by the Evaluation Management team (EMT), in cooperation with the chairs of the main and subject panels, and are hierarchized (*first and second reader*). Both are to prepare an evaluation report of an excellent output; the report consists of a grade (5 stars highest quality – outstanding, 1 star lowest quality – poor) and a narrative (100 to 200 words). It is the responsibility of the main referee to present a report with both individual reviews and an evaluation reaching a consensual grade. The referees are, a priori, informed of the definitions of the terms „*originality*“, „*significance*“ and „*rigour*“ within the field, based on the conclusions of the calibration exercise. The descriptions of the quality levels used for grading of individual excellent outputs are identical to the descriptions of the quality levels necessary for grading *Scientific research excellence* of the whole research unit RU. It is supposed that the referee processes 10 outputs daily, regardless of their size and form, and will receive a daily compensation of 500 €.

The KA4 team made a number changes to the original proposal, either for the purpose of speeding up the whole process out of necessity, due to limited time (choosing the referees April – May 2015, evaluating May–June 2015), or due to the methodology in cases when the recommendations of the TP were not considered adequate.

1. Due to the fact that scientific organisations such as IBRO were included in the pilot test, it was considered unacceptable to limit the amount of excellent outputs solely to scholarly outputs, and the option of reporting outputs of applied research that appear in RIV, the Czech R&DI Information system (patents, licenses, models, etc.) was also given („*non-scholarly outputs*“). Further, English was not strictly required, mainly due to the fact that it is not a dominant language in outputs in humanities and social sciences.
2. Referees were chosen solely by the members of KA4 team, as choosing the panel chairs, who were according to the TP proposal responsible for allocating the referees, occurred practically simultaneously. Also, calibration meeting was held in the second half of May, when part of the excellent outputs was already handed over to the referees and it was not possible to perform the instructions of referees, as recommended by TP. Besides, there was no need to interpret the three terms used („*originality*, *significance*, *rigour*“) specifically for the individual fields as was concluded in calibration meetings of all three disciplinary area panels.
3. The referees were not hierarchized, worked independently and the output of the evaluation of every excellent output was two equally valid grades, accompanied by narratives presented to the subject panel. The process of creating one consensual grade based on interactions between two referees is, from a logical standpoint, fairly complicated, extends the whole evaluation, and, in the KA4 teams' opinion, is conceptually wrong. This type of procedure would be in conflict with the established practice of peer-review, in which the individual referees (e.g. of scientific publications) are fully independent in their opinions.

4. The descriptions explaining how the grade is assigned to the individual excellent output were (based partially also on a contribution made by Daniel Münich from KA2 team) changed dramatically in comparison to the recommendations of the TP, as is clear from the Referees' Guide (see **Background document 6**, Appendix V). The referee is actually expressing views on, logically, only one excellent output, without the need to understand the context of the research unit where the publication was created and, therefore, does not generalise the evaluation in terms of the whole research unit, as the descriptions of the TP confusingly implied. To compare, we quote the beginning of a description used for the highest grade of output "outstanding": „*the output is internationally outstanding, meeting the highest standards*“, in contrast to the original: „*the RU is a global leader*“. Furthermore, the evaluation based on a number of stars (5 to 1) was replaced by grades A, B, C, D, E, similarly to grades in the evaluation of the whole Research Unit, in order to avoid any misunderstanding.
5. The template for referees was, in contrast to the original recommendation of the TP, simplified, though five basic bibliometric indicators (*citation number, category expected citation rate, citation percentile, journal impact factor, journal percentile*) were added for publications registered in WoS. The extent of the narrative was halved (50 to 100 words) for articles in impact journals and left unchanged in the extent of 100 to 200 words based on the TP recommendations for "large" outputs only (monographs and chapters), for which bibliometric data is usually unavailable (see **Background document 6**, Appendix V).
6. In terms of compensation, we took into consideration the fact that evaluation of publications in impact journals was dramatically less time consuming (the bibliometric data given serve as a guideline) than evaluation of large outputs, often several hundred pages long, which prevail in social sciences and humanities. Compensation of evaluation of the former was set at 50 € per one output (roughly the same sum as expected by the TP), whereas compensation for evaluation of the latter was doubled (100 €).

7.2 EVALUATING EXCELLENT OUTPUTS AND ORGANISATIONAL MATTERS

A total of 31 research units presented 235 excellent outputs, out of which 16 were in the category „non-scholarly“, see Table . An overwhelming majority of the „scholarly“ outputs were journal publications in English in science and engineering fields, available in PDF format, of which 90% were registered in the Web of Science. On the contrary, in humanities, these were mainly large books and chapters primarily in Czech, amongst the foreign languages English was not predominant, and none of the outputs presented were in the WoS. Parts of the books were available only in print and had to be transferred into PDF versions by the MEYS administrative team. Non-scholarly outputs were mainly proven technologies (9) and patents (7). Their descriptions (PDF documents) were, in some 4, 8 and 4 cases in English, Czech with an English abstract and, in Czech only, respectively. Therefore, it was decided that non-scholarly outputs would not be sent to independent referees, but would be evaluated by subject panels during meetings in Prague and with the help of Czech advisors (their places were provisionally taken by Czech members of the Main Panel).

The process of remote evaluation of excellent outputs consisted of three phases:

- obtaining and contracting qualified referees, willing to prepare evaluations of several excellent outputs in the area of their expertise within a short time frame (typically two to three weeks),
- ensuring on-line access of the referees to the outputs and, in return, obtaining grading and evaluations,
- converting the results of remote evaluation of individual outputs into integrated form, which would consequently simplify the work of the panel.

The KA4 team contacted internationally renowned experts from relevant scientific communities they are in touch with outside the Czech Republic, primarily middle-aged, working at universities as at least Associate Professors, and in scientific organisations, as independent or managing researchers. A specific personalised data folder was established in the on-line support system for each referee, with guidelines for referees (see Appendix V „Guidelines for referees“, **Background document 6**), PDF versions of the excellent outputs that were to be evaluated, and templates for filling in the evaluations in form of a grade and a narrative. The grades and narratives were then copied into integrated excellence reports on the level of Research Units from which, subsequently, an integrated excellence report on the whole field was created. This report contained information from every research unit on the total number of excellent outputs submitted and evaluated by independent referees, the number of outputs in which the grading of the two referees differed by more than two grades, and the total number of grades A to E given. An example of these two integrated reports in the form of concise tables is given in the Table 5 and Table 6.

Table 5 Example of integrated report on excellent outputs by one panel

Code RU	Type of RO	Scholarly outputs	Non scholarly outputs	Scholarly Excell. Outputs	Non scholarly EO	EO reviewed	EO rev. by 2 ref.	EOs score diff. 2 points or more	Number of score				
									(A)	(B)	(C)	(D)	(E)
AA_F_14	ScRO-HEI	542	72	8	1	8	2	1	X	X	X	X	X
BB_F_14	ScRO-HEI	433	56	10	0	10	2	0	X	X	X	X	X
CC_F_14	ScRO-HEI	619	23	8	0	8	1	0	X	X	X	X	X
AB_FCH_14	ScRO-HEI	319	200	9	0	9	4	1	X	X	X	X	X
BA_FCH_14	ScRO-ASCR	1142	5	20	0	20	9	0	X	X	X	X	X
AC_HP_14	ScRO-ASCR	146	7	3	0	3	0	0	X	X	X	X	X
CB_14	IBRO-RTO	119	90	2	3	5	0	0	X	X	X	X	X
CA_14	IBRO-RTO	21	21	2	1	2	1	1	X	X	X	X	X
AD_14	NatRes	28	3	2	0	2	0	0	X	X	X	X	X

Table 6 Example of integrated evaluation report on excellent outputs of one research unit RU

OECD Field	Type of output	Full reference	Citations	Exp. Cit. rate	IF	Percentile IF	Score of the first referee	Assesment 1	Score of the second	Assesment 2
Chemical sciences	Journal article	SETTER: web server for RNA structure compa	3	5,91	8,28	9,0%	X	The rapid inc		
Chemical sciences	Journal article	Selective Synthesis of 7-Substituted Purines via	12	9,93	6,14	7,8%	X	The 2- and/or		
Chemical sciences	Journal article	The influence of electrolyte composition on elec	14	12,15	1,84	76,6%	X	This paper de		
Chemical sciences	Journal article	Flavin-cyclodextrin conjugates as catalysts of e	31	15,83	6,38	31,6%	X	Enantiomeric		
Chemical sciences	Journal article	Unprecedented Meta-Substitution of Calixarene	19	4,54	6,14	7,8%	X	In the paper t		
Chemical sciences	Journal article	New insight into the role of a base in the mecha	6	1,41	3,81	25,3%	X	Enantiomeric		
Chemical sciences	Journal article	Experimental study of hydrocarbon structure ef	13	7,89	2,56	30,4%	X	Although the	X	The results o
Chemical sciences	Journal article	Correlation of oxidative and reductive dye blea	32	16,75	2,42	72,2%	X	The main goa	X	This work wa
Chemical sciences	Verified Technology	Technology of ammonia stripping with vapor re								

Bibliometric indicators are explained in section 6.1.

7.3 EVALUATION RESULTS

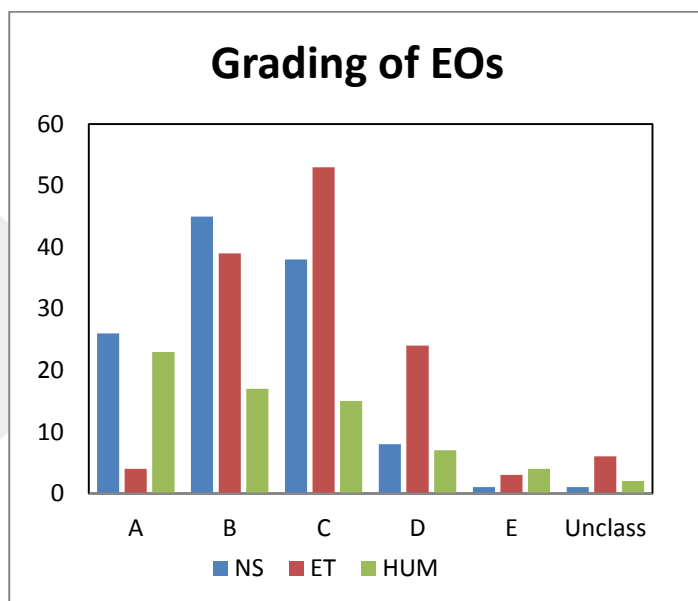
25 foreign experts took part in the evaluation. Due to time constraints, we were unable to secure qualified referees for all 219 scholarly outputs, only 184 outputs were evaluated, out of which one third (65) were evaluated twice. The number of outputs evaluated by one referee was between 1 and 25. Scholarly outputs that remained unevaluated (35) were classified by the subject panels during meetings in Prague, similarly to 18 non-scholarly outputs. In some panels, outputs where the second evaluation was missing were still evaluated, and the total number of outputs with two evaluations increased to 89. Whereas in sciences and engineering, the evaluations of roughly 10% differed by 2 grades, in humanities, this was 40% of the outputs, which shows a much larger degree of subjectivity in this disciplinary area in comparison to exact sciences. The table and bar graph (see Table 7 and Figure 2) show the total number of qualitative grades given in the five point grading system of evaluation; the significant differences in the distribution of grades within these three disciplinary areas are obvious.

Table 7 Overview of registered excellent outputs divided by disciplinary area and results of evaluation by independent referees

Disciplinary area	EO scholarly	EO non scholarly	WoS	Nr. indep. referees	Nr. graded schol. EO by referees	Nr. scores by referees	Nr. graded EO (refer. + panellists)	Nr. scores (refer. + panellists)	EO scores diff. by 2 or more
Natural Sciences	80	5	78	13	72	106	83	119	4
Engineering and Technology	80	11	71	13	63	89	91	129	4
Humanities	59	0	0	9	49	54	53	68	6

Disciplinary area	Number of score					
	A	B	C	D	E	Unclassified
Natural sciences	26	45	38	8	1	1
Engineering and Technology	4	39	53	24	3	6
Humanities	23	17	15	7	4	2

Figure 2 Qualitative grades given to excellent outputs in disciplinary areas



Note: NS Natural Sciences, ET Engineering and Technologies, HUM Humanities

Significant differences in exact sciences (Natural Sciences and Engineering and Technology) on the one hand, and Humanities on the other hand appeared in the process of choosing and evaluating excellent outputs, with regards to both the dominant type of excellent outputs (articles in English in international magazines vs. books in Czech mostly by local publishers), as well as the approach of the referees (a large degree of consensus vs. a wide range of evaluations). In exact sciences, outputs of small size registered in the WoS are predominant and bibliometric indicators are an important tool, therefore evaluating publication is not too time consuming. In humanities, this concerns extensive chapters or even monographs with hundreds of pages that require much deeper insight, combined with at least a passive knowledge of Czech of the referee, and the evaluation can therefore be much more time consuming. Moreover, in this disciplinary area, the opinions of referees can be rather subjective. Using international reviews of books (as Academy of Sciences of the Czech Republic has been attempting in the first phase of its evaluation of its teams in 2015) could help make the evaluation process more objective. Another issue is evaluation of non-scholarly outputs, which are also mostly in Czech.

The K4 team asked the referees for feedback in form of a short, 5 question questionnaire, slightly differing for disciplines in sciences and humanities; the responses were sent by 13 (former) and 5 (latter) referees. Both groups regarded evaluation of excellent outputs as useful and most referees in exact sciences assessed the use of bibliometric indicators as positive. The referees in sciences considered the descriptions of the five level grading system clear, however three out of five referees in humanities did not. The compensation for work performed was considered adequate by most, though almost all agreed on the fact that they had spent more time evaluating the outputs than the KA4 team had expected (roughly up to one hour for “small” outputs and two hours for “large” outputs). This implies that the referees’ compensation does not have crucial impact on the thoroughness of the evaluation. The KA4 team also met with refusal to participate as a referee in the pilot evaluation, mainly due to time constraints. Though in a few cases, specifically in exact sciences, this was due to methodological disagreement with secondary evaluation of scientific work and the belief that the quality of the output itself is already proven by the journal impact factor and citation index, both under the standardisation of the given subject.

In most comments (80%), the panel members assessed the evaluation of chosen outputs as a positive characteristic of the new Methodology; however, opinions were not unified on the issue whether this work should be performed solely by independent referees and what the role of the subject panel is, in this respect. Also, proposals were made to integrate the outputs chosen in a better way into the context of scientific work of individual scientists or teams within the Research Unit and, consequently, improve the degree to which excellent outputs can be personalised. For more information on feedback from the panel members, see 10.1.2.5.

7.3.1 Recommendations for Evaluating Excellent Outputs

1. Allow registration of non-academic outputs (e.g. patents, technologies).
2. Both reviewers work independently, they will prepare two independent reviews of an excellent output.
3. The explanatory notes on assigning the grades will be based on the suggestion used in the pilot test (see the **Background Document 6**, appendix V).
4. Bibliometric indicators of outputs will be added to the referees' template, wherever possible and relevant.

8. The work of the panels

8.1 REMOTE EVALUATION

The basic methodological document for the members of the main and subject evaluation panels was the “*Expert Panel Member Guidelines for the Pilot Testing*”, see **Background document 6**). The document contained a description of the organisation of R&D&I in the Czech Republic and the basic features of the proposed methodology of evaluation, a description of the pilot test, a description of the documents available to the panel members (the Research Unit self-assessment report, the Research Unit bibliometric report, a summary report on excellent outputs), instructions for the evaluation and instructions for working with the support system (see Chap. 2). Furthermore, the panel members were given a document prepared by TP - „*First Interim Report: the R&D Evaluation Methodology*”, tables with overviews of the number of workers, numbers of outputs and relevant bibliometric indicators of the Evaluated Unit and their Research Units, prepared by the KA4 team (see **Background document 7**), as well as Submission Guidelines – guidelines for filling in the Research Unit and Evaluated Unit self-assessment reports (see **Background document 1**).

Due to lack of time, it was not possible to hold an informative and, especially, a calibration meeting of all the individual subject panels before launching the remote evaluation. After calibration meetings in May (see Chap. 5), the subject panel chairs were requested to perform calibration in their panels in remote form. The result of the simplified calibration was unsatisfactory, mainly due to a lack of clarity in the content and, therefore, in the outputs of this calibration meeting (see also Chap. 5 and 11).

Remote evaluation was launched in the middle of May 2015, after the PDF version of the registered Research Units’ self-assessment report was prepared and made accessible in the on-line support system to both the subject and main panel members. The results of the subject panels’ remote evaluation were supposed to be available in the on-line support system prior to launching the personal panel meetings in Prague, i.e. by June 29, or by June 30, 2015 for the Natural Sciences, and Engineering and Technology panels, resp., and by July 6, 2015 for the Humanities panel.

The chairs of some of the subject panels, who evaluated a larger number of research units, allocated a lead and a second evaluator to each of the research unit, whose role was to prepare an evaluation proposal for the personal panel meeting in Prague.

The members of the panel had access to an evaluation form and a summary report on excellent outputs of the given panel in the folder of the relevant panel in the on-line support system, as well as documents from the individual Research Units in the individual subfolders: the self-assessment and bibliometric reports and PDF files with excellent outputs.

8.1.1 Recommendations for remote evaluation

1. Remote evaluation must be preceded by a calibration meeting of the panel members of the individual subject panels (see Final Report 3: The Small Pilot Evaluation and the Use of the RD&I Information System for Evaluation, section 2.3.3).

2. The settings of the evaluation form in the information system should enable a member of the administrative team or a member of the main panel to observe timely and full completion of the form. An auxiliary routine should provide generating and sending email alerts.

8.2 PERSONAL PANEL MEETINGS IN PRAGUE

8.2.1 Logistics

The personal panel meetings were held in Prague during two weeks from June 29 to July 3, 2015 for the main and subject panels in two disciplinary areas, Natural Sciences, and Engineering and Technology, and from July 6 to July 9, 2015 for the main and subject panels in the disciplinary areas of Humanities in the building of the National Technical Library. The preparation and course of the personal meetings were ensured by the KA4 team members, together with effective support provided by the administrative team of IPN Methodology and the panel secretaries.

Every panel was provided with a trained secretary, whose main task was to ensure administrative support to the panel members, ensure the organisation of the meetings, journeys of the foreign experts and their stays in the Czech Republic, and participate in the meetings of panels in cooperation with the administrative team of IPN Methodology.

The foreign experts of the evaluation panels were contracted as experts and “Agreements to Perform Work” were signed with them. This brought about a number of complications, including the need of the panel members to purchase and pay for their own plane tickets to attend meetings in Prague, payment of compensation for work performed after filling in the timesheets, reimbursement of personal expenses with a several month long delay, all of which, in effect, led to justifiable criticism expressed by the panel members (see comments, section 10.1.2.9).

8.2.2 On-site visits

The proposal of the Methodology as suggested by the TP does not take into account on-site visits. The background material, i.e. the self-assessment and bibliometric reports, and evaluation of the excellent outputs should be satisfactory for the expert panel members to carry out evaluation. Not until the conclusion of the small pilot evaluation⁷, published by the TP in May 2015, did it state that the panel members had in majority agreed that *“on-site visits are useful particularly for organisations that have responded to questions in the self-assessment report insufficiently or have misunderstood the questions. In the case of the latter, the visit helped clarify the questions that the RU answered incorrectly. The panel members recommended that if on-site visits were to be completely excluded due to high costs, the evaluation agency should consider organising Q&A sessions via on-line tools or videoconferences.”*

During the preparation phase of the pilot test we decided that due to the time constraints we would realise on-site visits to several chosen workplaces with the objective of assessing their importance and necessity. An overview of the on-site visits carried out is given in Table 8. It was not possible to realise the visit of the Chemical Engineering panel (2.4 Chemical Engineering), due to categorical refusal of accepting the panel members’ visit by one Evaluated Unit.

⁷ See Background report 10: The Small Pilot Evaluation – Feedback and Results, section 2.2.1, prepared by the Technopolis Group and its subcontractors

Comments of panel members and Evaluated Units on on-site visits are mentioned in Chapter 10 (see sections 10.1.2.4 and 10.2.1.10, particularly).

Table 8 On-site visits

1. Natural Sciences				
PANEL	PANEL MEMBER	RESEARCH ORGANISATION	RU	DATE
1.6 Biological Sciences	Arnold JM Driessen +2	Faculty of Chemistry, Brno University of Technology	VUT_FCH_16	30.6.
1.4 Chemical Sciences	Michael Haines+1	Research Institute of Inorganic Chemistry	VUANCH_14	1.7
	Philippe Hapiot+3 Erik Thulstrup	J. Heyrovský Institute of Physical Chemistry of the AS CR	AVCR_UFCH_14	1.7
2. Engineering and Technology				
PANEL	PANEL MEMBER	RESEARCH ORGANISATION	RU	DATE
2.5 Materials Engineering	Gijsbertus de With+3 Jonathan Seville	Institute of Chemical Process Fundamentals of the AS CR	AVCR_UCHP_25	1.7.
2.7 Environmental Engineering	Wim H. Rulkens +2	University of Chemistry and Technology	VSCHT_FTOP_27	1.7.
2.9 Industrial Biotechnology	Fons Voragen+1	University of Chemistry and Technology	VSCHT_FPBT_29	1.7.
6. Humanities				
PANEL	PANEL MEMBER	RESEARCH ORGANISATION	RU	DATE
6.1 History and Archaeology	Leoš Müller+1	Faculty of Arts and Philosophy, University of Pardubice	UP_FF_61	8.7.
	Frank Hadler+2	Institute of Contemporary History of the AS CR	AVCR_USD_61	8.7.
	Frank Hadler+2	National Archive	NA_61	8.7.
6.2 Languages and Literatures	Guy Achard-Bayle+3 Michael North	Faculty of Arts, University of South Bohemia in České Budějovice	JU_FF_62	8.7.
6.3 Philosophy, Ethics and Religion	Jakob De Roover+2	Faculty of Arts and Philosophy, University of Pardubice	UP_FF_63	8.7.

8.2.3 Recommendations for on-site visits

1. Perform an analysis of importance and usefulness of the on-site visits, considering the objective of the evaluation. Should a formative characteristic of the evaluation be accentuated, the evaluation cannot do without on-site visits.
2. If the on-site visits are realised, it will be necessary to prepare the logistics and consider the increased costs of evaluation.
3. As a compromise, particularly for evaluation in the disciplinary area of social sciences and humanities, a form of joint meetings of the panel members and representatives of the evaluated unit held at the location of the panel meetings could be considered.

8.2.4 Recommendations for the organisation of on-site visits

1. To prepare detailed scenarios and instructions for both the panel members and all the participants in the research unit RU (including a proposal of visitation of the laboratories, libraries, etc.).
2. The panel members should prepare a set of questions and provide them to the research unit prior to their visit.
3. Appoint persons responsible for moderating the dialogues.
4. Prepare the recommended content and extent of the presentation focusing on introducing R&D activities (prevent repetition of facts already included in the self-assessment report).

8.2.5 Preparation of evaluation reports for the Research Units and Evaluated Units

8.2.5.1 *Evaluation reports for Research Units*

During personal meetings of the subject panels, the results of remote evaluation of the individual Research Units were discussed and evaluation in the form of a grade and narrative evaluation were consensually prepared. The final edit of the evaluation report for the Research Unit was then, mostly after the personal meeting took place, prepared by the panel chair. The main panel chair was responsible for approval of the evaluation report; the main panel chair also often expressed comments, which were then incorporated in the report.

Some subject panels that evaluated a larger number of Research Units followed the recommendations of the TP. A lead evaluator, supported by another panel member, was responsible for the preparation of the proposal of the remote evaluation.

The results of Research Unit evaluation, i.e. both the grade and the narrative, proved to be substantially systematically different in the evaluation of individual subject panels and were criticised in the comments of the Research and Evaluated Units. Furthermore, some of the Research Units have also commented on the lack of sophistication of the accompanying narrative comments. All of this was caused by insufficient calibration, which was performed, due to time constraints, only in one common meeting of the members of the main and chairs or deputy chairs of the subject panels.

The template of the Research Unit evaluation report is part of the Expert Panel Member Guidelines (see **Background document 6**, appendix VI). In the bar graph, (see Figure 3 to Figure 7) the total numbers of grades given in all 5 criteria, divided according to the disciplinary area, are shown.

Figure 3 The number of qualitative grades based on criterion I *Research environment* given in the pilot test

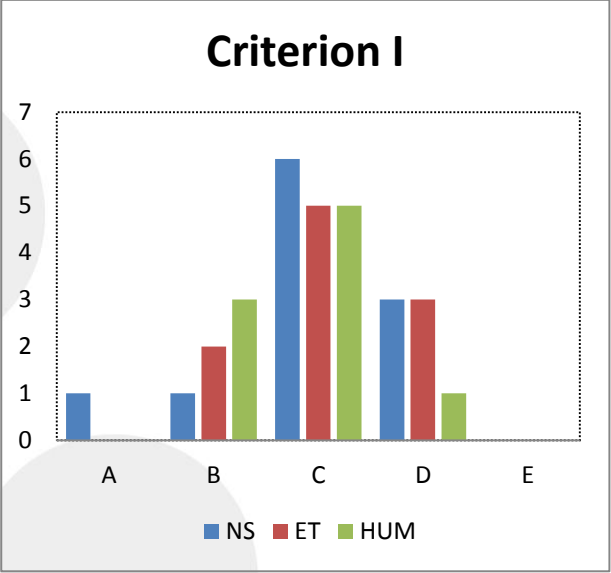


Figure 4 The number of qualitative grades based on criterion II *Membership of the global and national research community* given in the pilot test

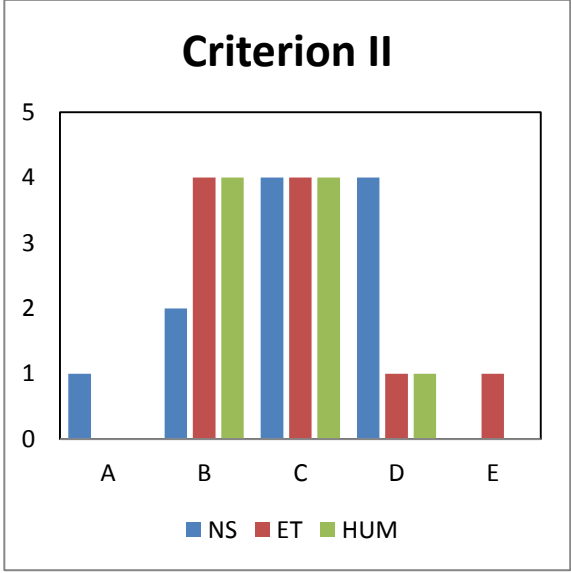


Figure 5 The number of qualitative grades based on criterion III *Scientific research excellence* given in the pilot test

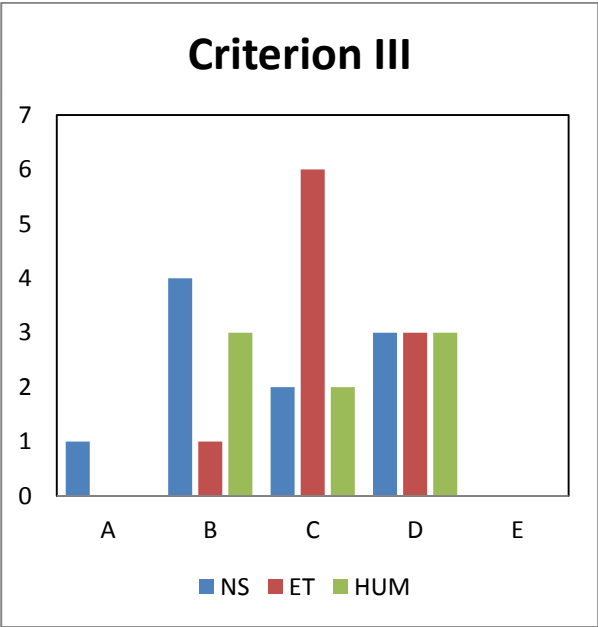


Figure 6 The number of qualitative grades based on criterion IV *Overall research performance* given in the pilot test

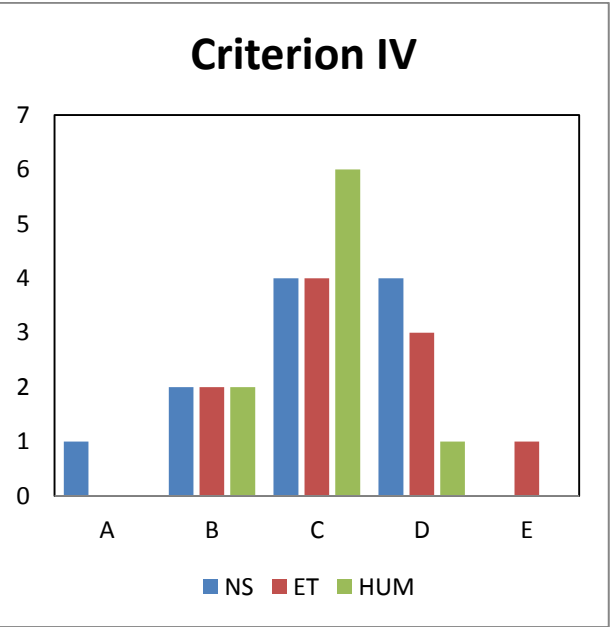
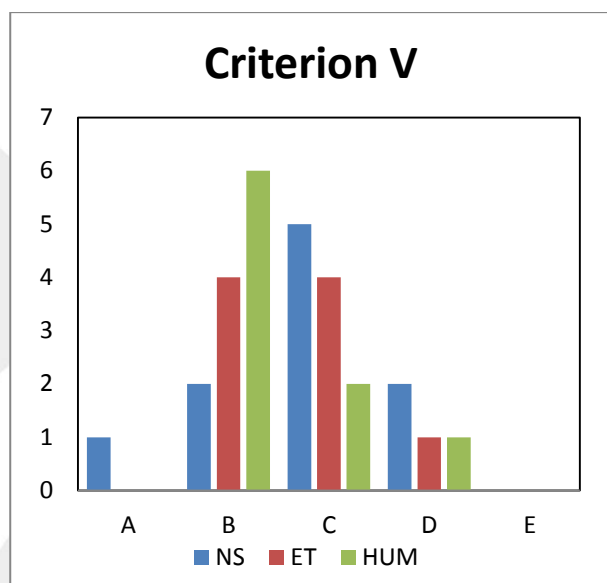


Figure 7 The number of qualitative grades based on criterion V *Relevance for society* given in the pilot test



8.2.5.2 Recommendations for Research Unit evaluation

1. It is necessary to leave more time for the calibration meetings of the main panel, as well as inside the subject panel, calibration meetings should take place firstly before the remote evaluation is performed, then again before the personal panel meetings are held and later, when the panel meetings take place.
2. It is necessary to ensure that there is enough time to prepare most of the evaluation report during the personal meetings.
3. The information system (IS) should ensure that the individual edits of outputs of the individual panel members are archived and that the final version of the report is provably approved in the IS by all the panel members.

8.2.5.3 Evaluated Unit evaluation report

Preparing the EvU evaluation report was negatively impacted by several factors. Firstly, the instructions of the TP (see Summary report section 3.2.4: „*The Subject Panel Chair for the major field of involvement of an EvU will be responsible for the analytical report at the EvU level, based upon a draft developed by the Evaluation Management Team and in cooperation with the other relevant Subject Panel Chairs*“) was methodologically ambiguous. The KA4 team therefore recommended that the EvU evaluation report be prepared by the main panel chair, in cooperation with the chairs of the relevant subject panels. Secondly, the members of the expert panels criticised the insufficiency of the background materials, particularly the self-assessment reports focusing on primarily the level of Research Unit, used for the preparation of the summary report of the Evaluated Unit that is composed of several Research Units. Thirdly, during the personal meetings every panel chair or deputy chair had approx. 1 to 1.5 hours for each EvU, which proved to be inadequate. All this impacted both the quality of the evaluation reports prepared and the disproportionately long time needed for their completion negatively.

8.2.5.4 *Recommendation for Evaluated Unit evaluation*

1. It is necessary to decide if the EvU evaluation report is truly necessary and brings added value that is proportionate to the considerable efforts, time and costs needed for its completion.
2. Enough time must be made for EvU calibration meetings before realisation of the evaluation at workplaces, as well as the time for the meetings of the representatives of the individual research units RU, that form the evaluated unit EvU.
3. It is necessary to ensure that all the representatives of the individual research units RU that form the evaluated unit EvU have included their inputs on the evaluation of the EvU in the information system in such a way that the final evaluation may be processed by the main panel chair. The functionality of the information system will also enable provable approval of the final version by the panel representatives.

8.2.6 **Field of science and disciplinary area evaluation reports**

The TP recommended (see Summary report section 3.2.4) preparing analytical reports on the field of science and disciplinary area levels at the end of the evaluation process, under the supervision of the subject panel chair and main panel chair. Only a chosen sample of Research Organisations and the Research Units registered by them took part in the pilot test, therefore these above mentioned analytical reports were not prepared due to lack of coverage of the disciplinary area and fields of science.

9. Compensation

In its background material, the TP⁸ proposed compensation for experts taking part in the evaluation, as shown in the table below (see Table 9).

The KA4 team carried out its own research of remuneration typical for evaluating outputs in R&D&I abroad, considered the short time span in which the foreign experts had to make a decision, and the dates during which they were supposed to work, as well as taking into account the possibilities given by the overall budget of the IPN Methodology budget. Keeping this in mind, they proposed the remuneration sum for foreign experts, panel members and referees for the pilot evaluation as given in the Table 9.

Table 9 Daily remuneration of experts

	Daily fee of TP	Daily fee during the pilot test
Foreign Experts		
Main and Subject Panel Chairs	€ 1,000	€ 500
Subject Panel member	€ 800	€ 450
Consultant Specialist	€ 800	positions were not filled
Referee	€ 500	€ 500
Czech Experts		
Main Panel Member	€ 500	CZK 3,200
Consultant Specialist	€ 500	positions were not filled

9.1 PANEL MEMBERS

The remuneration of panel members was stipulated differently in comparison to the proposal of TP. The Ministry of Education, Youth and Sports stated that foreign experts in the role of Main Panel Chair and Subject Panel Chairs would receive a gross hourly compensation of CZK 1,750, and these in the role of the Subject Panel members would receive a gross hourly compensation of CZK 1,575.

9.2 REFEREES

When considering the compensation for the referees, the basic thesis of the TP stated that the referee will evaluate 10 articles in a day and that the compensation for one article in the disciplinary area of Natural Science and Engineering and Technology would be € 50. In the questionnaire given to the referees after the evaluation we asked, among other things, if the remuneration could be considered generous (1 positive answer), adequate (10 positive answers) or inadequate (no positive answer). In the disciplinary area of the Humanities, where the prevailing outputs are books (monographs) or chapters in a book, the fee given was € 100 per output. None of the respondents considered it generous, 4 evaluated it as adequate and one respondent described it as inadequate.

⁸ Background report 4: Detailed evaluation cost framework

The type of contract, i.e. “An Agreement to Perform Work” that had been concluded between MEYS and the foreign experts was generally an issue for both groups of foreign experts, i.e. panellists and referees. This type of contract poses certain administrative problems for both the employee and the employer, firstly due to social and health insurance tax duties in cases when monthly compensation exceeds CZK 10,000, secondly, concerning the complications associated with complex and non-unified legislation stipulating income tax of individuals in the particular countries.

9.3 RESEARCH/EVALUATED UNIT

Due to the fact that participation of the Evaluated Unit was voluntary and that their employees performed the work connected with the project outside the scope of their normal work duties, based on the recommendation of the IPN Methodology team, Agreements to Perform Work for the purpose of preparing the self-assessment report and other work associated with the pilot test were concluded with specific workers that were nominated by the RU/EvU units. The number of workers, whose participation in the project was enabled by the specific RU/EvU this way, was based on the number of outputs of the given unit. The hourly rate was CZK 350. The total sum paid out to the workers participating was between CZK 50,000 to 200,000 per one EvU.

The Table 10 shows some direct and indirect costs connected with the work of external workers participating in the project of the pilot test, except the IPN Methodology team members. Only compensation provided to the RU/EvU units for preparing the self-assessment reports and for other activities performed as a part of the pilot test is included in indirect costs.

Table 10 Direct and indirect costs of the pilot evaluation

Position	Number	Total sum in thousands of CZK
Panel Chair (foreign)	12	4,807
Panel member (foreign)	23	5,638
Main Panel member (Czech)	6	286
Travel expenses of the panel members, including accommodation	35	608
Referee	25	711
Experts of research organisation participating in the pilot test	54	1,959
Panel Secretary	12	310
Total	132	14,319

Note: The full sum includes gross remuneration and all obligatory tax payments, with the exception of travel and accommodation expenses.

9.3.1 Recommendations for compensation

1. When formulating the rules of the next (full) evaluation, it is necessary to consider the different concepts of stating remuneration, as in a relatively large number of (European) countries, remuneration is determined after tax. It would therefore be appropriate to have Czech tax and legal experts investigate the issue and consider the impact of taxation in various countries of origin of the experts when determining the compensation, in order to avoid situations when the taxation of the experts in their countries of origin is inadequately high. Particularly agreements preventing double taxation should be pointed out, in which taxation in the country of residence is preferred, i.e. in their home country. Double taxation agreements also enable earning income as a freelancer.

2. Furthermore, formulating a type of work agreement with the foreign experts that would enable prompt operative payment of the experts' travel and accommodation costs during their work in the Czech Republic is necessary; furthermore, we consider the option of providing the expenses for this purpose in form of an advance or a flat rate prior to the trip to the Czech Republic to be appropriate. In case the above mentioned recommendations do not have any support in current legislation, we consider the importance of R&D&I evaluation in the Czech Republic important enough to create the specific legislation necessary to create non-bureaucratic access to foreign experts. Inconclusively, however, we believe that a possible type is the contract under the Civil Code §1746, which allows to solve both of the above named issues, i.e. taxation and flexibility when accounting and reimbursing travel expenses.
3. For the research organisations participating in the R&D&I evaluation, on the contrary, we recommend that all costs (including any compensation) is carried by the individual organizations themselves. The results of the evaluation, for a substantial part of the organisations, will be institutional subsidies from the state budget to support the organisation's operations, in other words the source of possible compensation for the employees who are involved in the preparation of the of evaluation.

10. Feedback on methodology

10.1 PANEL MEMBERS

All members of the main and subject panels were asked to answer questions related to the proposed evaluation methodology, as well as the process of pilot test. Some of the panels sent prepared summaries of their answers on behalf of the whole panel, others sent answers of individual panel members. Altogether, we received 26 sets of answers and comments from 12 main and subject panels, with a total of 41 panel members. You can find those in the separate document (see **Background document 8**).

10.1.1 General questions

In the first three questions, we focused on a general, overall assessment of the proposed Methodology and we asked about the positive and negative characteristics of the Methodology and for comments regarding a potential future Methodology.

10.1.1.1 Q-A. What, in your view, are the most important positive features of the proposed new R&D Evaluation Methodology and Funding Principles ("Methodology")? Why?

In response to this question, we received a total of 23 arguments in favour of supporting the Methodology; some of them appear regardless of their field specification: writing up a self-assessment report can be a useful tool for both the institutions evaluated as a whole, as well as for the purpose of identifying the importance of individual scientists for the development of individual Research Units. The fact that the main part of the evaluation is performed by foreign experts without any direct interests in the evaluated institutions guarantees, firstly, a high level of objective evaluation and, secondly, enables comparing evaluated Research Units in a wider, international context.

- a) In the case of exact sciences (Natural Sciences and Engineering and Technology), it is pointed out that similar, regular evaluation processes in these fields are commonplace as a part of evaluating science and research in developed countries. Therefore, this Methodology can facilitate classification of the evaluated research unit into generally comprehensible qualitative categories, which clearly define the research unit's importance within the international field's community, however, keeping in mind that the data available to the evaluators will be processed carefully and precisely, in a unified manner and based on the given directives.
- b) In the case of social sciences and humanities (SSH), it is emphasised that the Methodology, based on proven standards from other European countries, is going to set a clear set of evaluation rules that have so far not existed in the Czech Republic. Independent evaluation, performed by foreign experts, will enable easier application of positive experiences from abroad, together with evaluation of research units in SSH and will markedly decrease potential partiality, which stems from a limited circle of evaluators from our own country that are qualified and experienced enough to perform this type of evaluation. Applying Methodology of this kind will, first of all, support the activity of traditional, excellent science centres that have attained an important position in international research in SSH fields and, furthermore, will enable a more objective evaluation of the science-research potential in the case of newly established institutions. The Methodology has a significant potential as it enables going in the direction of field specific assessment, for which the need is given by a substantially different nature of the main outputs of

science and research in the fields of SSH, compared with natural sciences, and also the different language standards, since in a substantial part of the fields in SSH English is not the dominant language of the major basic research publication outputs (unlike fields in exact sciences).

10.1.1.2 Q-B. Please describe the negative features of the Methodology.

We have received a total of 25 critical arguments in response to this question, partially overlapping in content. Most of them do not address the actual principle of the Methodology itself, but more or less serious, specific technical solutions that the panel members have assessed, based on their experiences, as illogical or incorrect. The panel members generally recommend specifying some definitions, providing more time for panel meetings and, above all, for calibration meetings, which can enable eliminating a possible adverse shift in the scope of evaluation. It would be advisable to narrow the grades particularly between “B” and “C” (i.e. “+B”; „B“ a „B-“, etc.). Comments regarding the possible danger of subjective distortion of the evaluation results that could arise due to inappropriately chosen panel members (inaccurate field specialisation, not enough experience with scientific work, etc.) appear across all fields. Unlike subject panels (whose role in the evaluation process is clear), there appeared to be a certain level of confusion regarding the role of main panels and their panel members in the whole system. Further, the panel members pointed out the unclear assignment, i.e. they were lacking information on the way the results of the whole evaluation process would be subsequently applied, especially in the case of evaluating institutions of a completely different type. This objection was recurring, although it was stressed during the whole duration of the evaluation that it was a pilot test with the main objective to check the functionality of the proposed Methodology, not to evaluate a few chosen institutions.

- a) In exact sciences, the definition of a research unit (RU) was questioned. This definition does not enable evaluation to be performed within the logical, already existing and, to some extent, mutually independent research teams. Within the illogically constructed RU, we are unable to differentiate the share of individual teams on specific outputs that have been rated as excellent. The self-assessment report requires too much information that is subjective. It would be more suitable to limit the required information to objectively verifiable data, i.e. to divide this material into two parts (first – larger – with verifiable data; second – smaller – more concise – in written form).
- b) In the case of SSH, the panel members pointed out that despite individual differences it was clear that the basic evaluation principles of Methodology were based on experiences from evaluating natural sciences. Therefore, in spite of certain positive changes, some well-known weaknesses of the Anglo-Saxon, i.e. British model still remain in the Methodology. The evaluation grades for SSH are defined illogically; in these fields that are, to a certain degree, dependent on their local cultural and language environment, the „A“ grade („global leader“) makes no practical sense. The above mentioned danger of subjective bias of the results due to inappropriate selection of the panel members is even emphasised in SSH in the requirement of knowledge of the local environment (Central and Eastern Europe) and of the language skills needed. Without these, evaluation of the research unit, whose excellent outputs are in another language than English, is pointless.

10.1.1.3 Q-C. Please provide comments and/or suggestions you would like to add regarding the future of the Methodology.

The topics in the responses to this general question were, in part, repeating the same issues already present in question Q-B. The recommendation to also include the RU's idea of its future development in some form into the evaluation

process repeats across all fields. Furthermore, the question of the relationship between scientific research and education, in this case mainly on the level of PhD education, should be considered. All subject groups recommend providing the panels with a comprehensive list of publication activities (not only a list of excellent outputs), an overview of scientific organisational activities (organizing scientific congresses; publishing significant scientific journals, etc.) and a list of the individual members of the research units, including specification of their individual activities in publishing and organizational segments of research and science.

- a) As far as exact sciences are concerned, it is further recommended to define the relationship between RUs and EvUs in more detail, as the general SWOT analysis makes it difficult to abstract the positives and negatives of the individual RUs. During evaluation of the RU, clear statistical data concerning human resources in relation to the achieved outputs („output/FTE“) and financial resources obtained from competitive (targeted) funding should be available. Again, a request for a shorter text section in the individual criteria in favour of statistically accurate data that can be evaluated appears. Due to the demanding nature of the evaluation, the panel members recommend performing the evaluation process in a 5-year cycle, within which 20% of the RUs would be evaluated every year.
- b) Alongside the above mentioned common topics, the SSH panel members also recommend adding a short introduction to the self-assessment report, in which every RU would include a short overview of its development so far. In SSH subjects (especially “Humanities”), the panel members recommend mandatory representation of a field specialist with Czech language skills in every subject panel, taking into consideration the primary role of Czech language as a basic means of communication in the fields that are connected with the local cultural and historical environment due to their specialisation. The selection of the panel members must be done very carefully, due to the rapid innovative and methodological development in SSH fields that reflect social and cultural changes in the Central and Eastern European region in the last decades.

10.1.2 Specific questions

Nine more questions that the panel members were asked were related to a detailed, specifically defined characteristic of the Methodology.

10.1.2.1 *Your comments on the evaluation of the RU level vs. EvU level*

The Methodology, as proposed by the TP, has been built on the research unit as the basic unit of evaluation from its very beginning. Only after the comments of the IPN Methodology project team have the processes of evaluation on the level of EvU unit also been included. However, this proposal has a number of weaknesses and we have a few substantial comments related to this issue.

The responses of the panel members represent a fairly wide spectrum of opinion, regardless of field specialisation. Most responses to this question prefer the original proposal, based on which the basic evaluation unit is represented by the research unit (RU), however with certain objections to the definition of the RU (mostly this concerns the possibility of setting up more RUs within one field). The main objection to the efficiency of evaluation on the EvU level is the fact that EvUs are composed of, in part, by singular, individually evaluated RUs (which do not always have elements in common that can be evaluated), partly also by units that are not part of the evaluation. In this situation, it is very difficult to create an evaluation report of the EvU as a whole, if the EvU itself was not the subject of evaluation. In contrast, some panel

members are in favour of the evaluation of the EvU as a whole as they consider the RU to be too small of a unit (based on the type of the research unit evaluated).

10.1.2.2 Are the five assessment criteria relevant and properly formulated? Any suggestion for change?

More than half of the panel members responded that the evaluation criteria were proposed correctly and that they considered them to be optimal. Over 10% of the panel members recommended using a finer division than just five classification grades A to E, e.g. -, B+, B-, B-C, C-B. A couple of respondents had doubts about the usefulness of the classification grade A, “global leader”, taking into consideration that there are only few organisations in the world that can be evaluated this way. One respondent proposed that, on the one hand, the classification grades A and B should be combined into one, and, on the other hand, another grade should be added to the lower end of the scope, between D and E.

A few respondents from the field of Humanities pointed out that in specific subjects with national specialisation (e.g. Czech Literature, Czech Language, Czech History), the term “global leader” is inappropriate and should be rather understood as a national leader, cooperating with international institutions.

Some respondents highlighted that criteria numbers 3 and 4 are not independent.

A few respondents, particularly in the field group of Humanities, suggested a revision of the criteria that would let all the units evaluated understand the criteria during the process of preparing the self-assessment report in the same way and also provide comparable informative value for the evaluators. One respondent does not consider criterion number V “societal relevance” to be relevant for scientific organisations.

10.1.2.3 Formulation of questions and structuring of the self-assessment report according to the five criteria. Were all the questions (asked of the Research Units) useful? Would it be more advisable to limit the questions to a maximum of around ten and have an additional data-gathering section?

About half of the panel members who explicitly answered this question consider the formulation of questions in the self-assessment report to be appropriate. A few of the respondents recommended decreasing the number of questions and grouping them into two parts, one devoted to descriptive information (strategy, management, human resources, etc.), another containing data. Some of the panel members pointed out a lack of information, e.g. a full list of scientists and outputs (one respondent proposed the list to include hypertext links to a database source), more details on post-graduate studies, such as their duration, rate of completion.

The panel members also criticised the low standard of responses in the part of the self-assessment report devoted to scientific strategy of the research unit.

A few respondents criticised the overall quality of the self-assessment report. In summary of them all, we cite two: „*The self-assessment report is the principal material for the evaluation, but the quality of the reports presented by RUs vary a great deal. Note that since the present reports were prepared for this pilot exercise only, without any consequences for future financing, some units may not have paid sufficient attention to the report preparation*” and „*...some RU's were not*

totally aware that the “game” could be sort of valuable training for them, for next time when the evaluation is done for real.”

10.1.2.4 Do you consider on-site visits important or even indispensable? Would an interview of RU staff members by the panel in a neutral location be an option instead?

The opinions on the importance and usefulness of on-site visits oscillated between two different poles. A clear majority of the respondents (out of 24 explicit answers, 21 were completely or partially in favour of visits) considered on-site visits to be important and absolutely necessary, crucial to a successful evaluation. Also, a number of respondents recommended simplifying the procedure by inviting just a few representatives of the evaluated unit to the location where the panel meeting is held, as a number of respondents realised the logistical difficulties involved in providing on-site visits during a nationwide full evaluation of research organisations.

Several respondents described the on-site visits as very important and helpful to the evaluation, as only during the visits itself certain unclear issues relating to the self-assessment report were resolved. However, this is much more a sign of the evaluated RU's insufficiently completed self-assessment report rather than of a need and usefulness of the visit.

Three respondents doubt the usefulness of the on-site visit: one of them admitted to see a formative use in the visits, i.e. usefulness when compiling a recommendation for the future development of the unit. Another respondent suggested that the on-site visit, if carried out, should be realised on the EvU, not the RU level.

10.1.2.5 Do you consider grading the selected outputs by independent referees useful?

Questions number 5 and 6 (i.e. this one and the next) are closely related. In this section, we are going to analyse the responses only from the point of view of usefulness and helpfulness of obtaining evaluation of the given excellent output.

A clear majority of the respondents (out of 23 of the explicit responses, 18 were completely or partially positive) supported such a form of evaluation. Several recommended the comments on the evaluation included some key terms, such as originality, significance, rigour, readability/presentation, relevance to the field.

Two opponents of evaluation of excellent outputs by independent referees proposed that the evaluation is performed by members of the panels themselves. Here, it is important to point out that the opinions of the panel members on the evaluation of certain chosen outputs varied; some suggested the evaluation should be performed by the panel members, others have categorically rejected it.

The panel members from humanities requested they have a full list of all monographs published in the period evaluated available for the evaluation, furthermore accompanied by national and international reviews of the books.

A number of respondents pointed out the insufficiencies of the evaluation of excellent outputs. The panel members received the evaluations late, in case of some outputs it was impossible to obtain independent reviews, for several any reviews whatsoever. More on these insufficiencies, the reasons for them and suggestions on how to avoid them can be found in Chapter 7 “Evaluating Excellent Outputs” and Chapter 11 “Conclusion”.

10.1.2.6 Do you agree with assessment of only a small percent of research outputs or would you prefer to receive a full publication list? Would you be happy with a quality judgement based purely on bibliometrics, rather than an independent review of selected publications?

We received 9 explicit responses to this question, the other respondents mostly reacted in the previous question (number 5) or in answers to other questions. Three respondents prioritised the possibility to evaluate only a small number of outputs. Four respondents asked for a full list of outputs; however, we received a larger number of such requests for full lists of outputs, especially considering the responses to other questions. One respondent emphasised the importance of a list of types of outputs, i.e. not only scholarly (publications, books, etc.) outputs, but also non-scholarly ones, including outputs documenting cooperation with industries.

10.1.2.7 Do you consider bibliometric analysis important for your disciplinary area? What are your comments and/or suggestions to the content of the bibliometric report?

All the respondents from the Natural Sciences and Engineering and Technology field groups consider the bibliometric report useful. Some of them pointed out that the data in the bibliometric report must be evaluated keeping in mind the various publishing customs in different fields, and, as such, be a useful criterion in the evaluation.

Practically all the respondents representing humanities rejected the bibliometric report as unhelpful. One of the respondents considers the bibliometric report useful for articles published in scientific journals that are indexed in respected registers, such as Thomson Reuters or ERIH.

10.1.2.8 Do you consider a calibration exercise a necessary and useful part of the Methodology?

24 respondents explicitly answered this question, 22 responses were positive. Out of the two negative responses, one mentioned the impact of calibration would be limited and that in our version for the purpose of pilot test (see Chapter 5), the calibration had no meaningful role. These two negative statements prove an insufficient depth and width of calibration as a part of the pilot evaluation and emphasise the need for its improved definition in the next version of Methodology – see the recommendations in Chapter 5.

10.1.2.9 What are your comments on the pilot test (overall organization, materials submitted, panel meetings logistics)?

The reactions to the organisation of the pilot test were mostly positive. Several respondents expressed satisfaction with the website for sharing documents (see Chapter 2, On-line Support System). Some respondents criticised how late the documents had been sent and that they were too large. One respondent recommended that the panel members receive a “What to do” list that would help define and describe their tasks (see **Background document 8**, section “Other comments”).

Several respondents criticised the process of provision of plane tickets and accommodation, due to having to cover the costs themselves and being reimbursed later. Reimbursement took place several months later.

10.2 RESEARCH AND EVALUATED UNITS (RU/EVU)

The responsible or contact persons from the Research Unit were asked to respond to the questions regarding the proposed evaluation methodology, as well as the procedure of the pilot test. Most of the Research Units created a summary of responses on behalf of the whole Evaluated Unit, which did not impact processing or concluding from the comments obtained, and in a few cases submitted comments for the individual Research Units. Altogether, we received 20 sets of responses and comments from 31 Research Units. These are stated in the attached document, see **Background document 9**.

A summary of RU/EvU comments

- a) **obtained from Natural Sciences and Engineering and Technology** (12 RUs responded)
- b) **obtained from Humanities** (8 RUs responded)

10.2.1.1 Do you consider the structuring of the self-assessment report and formulation of questions to be answered as adequate? What was, in your opinion, missing and what was redundant or possibly leading to confusion?

- a) A critical tone prevails in the responses to this question, merely one research unit RU considered the self-assessment report ("SAR") to be unproblematic. Most comments recommend shortening the SAR, some even dramatically. Further, RUs require expanding the comments to questions in order to clarify exactly which aspects of the work of the RU will be evaluated and the criteria this evaluation will be based on, particularly a clearer definition of the tables with data. RUs from non-academic sector think that the structure of the SAR is suitable for academic research organisations. Most RUs also recommend not requiring a SWOT analysis. One RU pointed out the incompatibility of the time periods, for which some of the data and information was stated and another RU considered the evaluation period out-of-date, too distant from the present.
- b) Most RUs accepted the structure of the self-assessment report itself as adequate and user-friendly. However, certain objections (often identical in terms of content) to the usefulness of some parts of the report were repeated:
 - the definition of an FTE (50% of a full-time contract of individuals in the case of academic workers at public universities);
 - the point of determining the age structure is unclear, i.e. how are the data obtained going to affect the evaluation (in the case of social sciences and humanities the age limit of the presumed effective science output is set much higher than in natural sciences, technology and engineering);
 - the term "societal importance / impact" and the criteria it follows are not clearly defined.

10.2.1.2 Do you consider the formulation of five criteria as adequate; in particular, is it in your opinion useful to distinguish between the research excellence and research performance?

- a) From a total of twelve responses received, six tend to believe that the criteria are adequate, the other are partially critical, one is completely negative, one RU regards the criteria as unclear. Three critical comments point out that the criteria are suitable for the academic institutions, not for applied research organisations. One RU reminded that excellent outputs do not have to necessarily be in English. One RU regards the criterion „Membership in a national

and global research community” as misleading; as they believe that a sign of quality should be the organisation's ability to achieve excellence through its own workers. Apart from the criticism, one RU recommended taking inspiration from the criteria that are used by AS CR for its own evaluation. Whilst specifying the areas that are addressed, the questions that re-appear are almost identical to the questions that are contained in the self-assessment report.

b) Practically all RUs consider the proposed system of five criteria adequate; some without objections, others with certain comments. These comments refer to primarily two possible risks connected with using these five criteria under the conditions of SSH in the Czech Republic:

- an unclear definition of „excellence“, and criteria necessary to achieve this grade („global leader“);
- a concern that foreign experts, unfamiliar with the Czech environment, will be able to objectively assess activities of institutions, whose main science and research activities are primarily connected with the local Czech environment.

10.2.1.3 Is 1 to 2 per cent of outputs selected for grading reasonable for assessing excellence in research?

a) Three responses out of eleven agree with the fact that 1 to 2 per cent is adequate; one RU considers 15 to 20% adequate, 2 RUs require 5%, one RU 3%. One response states that it is unclear what excellence in research is. Further, RUs require not setting a maximum number (20 in the pilot test). Further alternative proposals include stating 5 to 10 outputs for every evaluated year, stating 3 to 4 outputs for all sub-fields (subjects) that the RU regards as dominant.

b) All RUs considered the number 1-2% (when setting an upper limit on the number of the outputs evaluated) questionable, but not for the same reasons. The objections were related to the size of the RU (for larger RUs, this number was acceptable, but not smaller RUs), as well as general structures of outputs in R&D in SSH. In these subject fields, there are generally more outputs than in natural sciences and the most important (excellent) outputs are often created over a period of many years (the main types of these outputs are books, often published in Czech). Choosing a small number of outputs in a short evaluation period can dramatically influence evaluation of the whole RU (both positively and negatively). These SSH specifics were considered also by subject panels, which always required (along with a few excellent outputs that were presented) additional submitting of a general overview of publishing activities of the whole RU in a longer time period than the period included in the self-assessment report (i.e. the timeframe given for choosing excellent outputs).

10.2.1.4 What is your opinion of the bibliometric report, do you consider the given indicators as easily understandable? What information is in your opinion missing and what information is redundant or possibly leading to confusion?

a) A significant majority (11 out of 12) have a critical opinion of bibliometric analysis. Several RUs state that bibliometric analysis makes no sense, that it is more of an interesting statistic rather than an important document used for evaluation, other RUs believe that it is incomprehensible. Some RUs request a much deeper and more inclusive bibliometric analysis. The rest of the RUs submit specific, detailed and extensive comments on individual indicators.

- b) RUs in SSH have a similar opinion to RUs in exact sciences, though the content and informational value of bibliometric reports, created as a part of the pilot test, are quite different in these fields. The submitted bibliometric reports are formally considered to be clear and concise; however their usefulness is, relative to the main point of the evaluation, disputable, because they contain only data about the numbers of, but not quality of the outputs. For SSH, the data referring to impacts in magazines or entries in WoS makes no sense, as these systems do not track the main (excellent) types of outputs in SSH fields at all - mainly books. Some RUs submitted proposals that specify which tools could be used in the future in order to contribute to the solution of this discrepancy.

10.2.1.5 How do you compare the proposed Methodology with the earlier version of the R&D evaluation (the “coffee grinder” and its recent modification)?

- a) From the 12 responses obtained, seven state that the current, valid version of Methodology is better than the proposed one, because it is more simple, accurate and objective. Two respondents believe that the methodology proposed is more objective, mainly due to the contribution of peer-reviews and a lower dependency on indicators, which are not objective and with incorrect weight. In two statements, which prefer the new proposal rather than the current state, it was emphasised that the proposal must be reviewed. Three RUs stated that they are unable to answer this question – two representatives of the RU wrote that they were not acquainted with the former methodology, the third said it is impossible to answer the question without knowledge of the financial impact. One answer reminds us that the evaluation is more suitable for academic institutions and puts ROs in applied research in a disadvantage.
- b) The opinions of the RUs varied greatly, depending on the features of their workplaces. RUs at the AS CR appreciate the proposed Methodology much more, which is logical considering they perform their internal 5-year evaluation in a similar way. In comparison, RUs at public universities prefer the current system of evaluation, which is based on conclusive and retroactively verifiable data (as logged in the RIV, R&D&I Information System). In the case of the proposed Methodology, they pointed out its dramatic subjective elements that impact the overall results in several phases (the self-assessment itself; subjectivity of the evaluators; lack of clarity in the definition of the evaluation criteria, etc.), complexity and how time-consuming it was. RU of different types (National Archive and National Museum) avoided answering the question directly and just repeated their opinion on the possible potential of the proposed system, without comparing it with the current one.

10.2.1.6 Are there any further comments and/or suggestions you would like to add regarding the future Methodology

- a) In the case of RUs in exact sciences, these comments are completely different in terms of their content, because they reflect the opinions of individual RUs. They exist on a wide range, from complete rejection of the proposed Methodology (recommendations to stay with the current system) to individual recommendations on improvements of the proposed Methodology. Mostly, the comments emphasise some of the negative (from the point of the view of the RU) elements that are contained and repeated in the Methodology. RUs in AS CR point out the quality of the current system of evaluation inside the AS CR, which they consider to be better than the proposal of the new Methodology. They recommend comparing individual RUs at AS CR and at public universities separately. In the case of other RUs, the danger of subjectivity in the evaluation is repeatedly emphasised (the level of quality of completion of the self-assessment report, the qualities of panel members and evaluators), as well as the fact that

the proposal of the Methodology gives an advantage to academic ROs over ROs in applied research. It is recommended to significantly simplify the evaluation criteria, limit the subjective human factor in the system, give more weight to objectively verifiable data (reduce the narrative part) and perform a “pre-selection” of RUs based on bibliometrics prior to the evaluation process itself. Less frequently, the RUs pointed out the positive potential of the proposed Methodology in their responses to the questions, i.e. its ambitions to build a system of evaluation based on research priorities of the state.

- b) In this point, most of the RUs in SSH reflected their general opinion of the proposed Methodology, already present in the responses to previous questions (see above 10.2.1.5). RUs in AS CR (as well as the National Archive and National Museum) do not express themselves in detail here, or just add individual proposals on how to improve the Methodology (i.e. bring it closer to the current system of evaluation, performed in AS CR). Specific comments refer to e.g. the boarder limit for individual RUs (50 outputs logged in the RIV), which is set unreasonably in the case of SSH (it does not take into account any kind of qualitative criterion). RUs at public universities repeat their negative opinion in different words and just explain in detail why this Methodology, in its current form, is unacceptable in practice for fields in SSH.

10.2.1.7 In your opinion, to what extent has the panel reached a correct view on performance in your research organisation?

- a) Most RUs (8 out of 12) stated that, in principal, it is possible to agree with the views of the panels and that the panels generally captured the position and results of the RU. A certain distance was expressed by two RUs (partial agreement, 85% match), disagreement (30% match) was expressed by one RU. Another RU stated that the strengths and weaknesses were captured by the panel completely accurately, but that it did not understand the strategy of the RU. Some RUs added certain objections to and comments regarding the compliance of the RU self-assessment with the evaluation of the panels. There are two objections that can be considered as major. Firstly, the absence of detailed and more expansive written feedback (mainly in connection with excellent outputs – the evaluation report is too short), and, secondly, a contradiction in the written evaluation and qualitative grades on a scale of A to E (the RU was evaluated better in written form than on the given scale).
- b) RUs in SSH overwhelmingly accepted the evaluation report as objective and fair. Partial objections related only to particular issues, such as insufficient justification of the assessment of a particular excellent output or insufficient data for bibliometric analysis.

In the case of all disciplinary area fields (exact sciences and humanities) two identical findings that need to be taken into consideration in preparing the final form of evaluation appeared repeatedly in the comments of the individual RUs (the conclusion of these views is a clear preference of a field structured system of evaluation, in which RUs that are field specific will be evaluated in one phase):

- The evaluation panel formed a realistic view of the RU only after the on-site visit.
- It is possible to evaluate how adequate the evaluation results are only within a framework of a larger number of field specific RUs, evaluated in the same time period. As a part of the scientific field community, a general and empirically proven awareness of the quality of the scientific work of one or another organisation exists. Structured results of any evaluation process refine such results and put the evaluated RUs from the same field on a qualitatively defined scale. In case the evaluation results show completely unexpected and inexplicable grades of

certain RUs on a qualitative scale (whether in a negative or positive sense), then the whole system may be questioned (even if this is due to a small error in calibration).

10.2.1.8 In your opinion, to what extent are the panel's conclusions and recommendations useful for the management of your institution?

- a) Half of the respondents (6 out of 12) believe that the conclusions and recommendations of the panels are useful to the management of the institutions (fully or in part) and that in the future they will, after critical assessment, work with them in some form. Six RUs believe that the views of the panels did not bring any new information or so far unknown conclusions to the management of the institution. Two institutions stated that they have their own evaluation that is better.
- b) All RUs consider the conclusions and recommendations more or less useful and helpful, though they might disagree with them.

10.2.1.9 Please provide an estimate of time and HR investment in preparing submissions for the pilot test. How many working days and how many people (in FTE) were involved (please provide an estimate for the administrative staff and for researchers)?

The answers to this question are quite indefinite, further they are cited as per the RU, EvU or RO. They vary significantly regardless of the field, generally we can say that the RUs in SSH stated much less of a time burden than RUs in natural sciences and engineering and technology. Based on data obtained in this manner, it is impossible to state some kind of "standard" burden, as the given time spent is very different even in RUs of a similar size (it ranges from several hours to hundreds of hours). If the RU was responding to the question of numbers of hours, contractually paid as a part of pilot evaluation by MEYS, then this number was assessed as suitable (though it did not cover all activities necessary to ensure the pilot test within the RU). Some RUs admitted that they did not devote the same attention (and time) necessary for a evaluation with real (financial) impact. It is clear that the evaluated institutions regard the realisation of Methodology as currently proposed to be very time consuming (if they state the number of hours spent in hundreds, including scientific workers in management positions, then such costs on the side of the institutions must be truly high).

10.2.1.10 Only for RU-s that received on-site visits: Please provide your comments on the merit, usefulness, and agenda of the visit (up to 250 words).

- a) A total of five subjects responded to this question (2 EvUs and 3 RUs) – three of which assess on-site visits mostly negatively – the visits occurred in inappropriate times, the faculty was evaluated by the wrong panel (however the faculty chose this panel itself by registering its RU), the visit brought on additional administrative and time burdens, the visit did not bring any further additional information to the EvU. One EvU and one RU assesses the usefulness of the on-site visit positively (the RU very positively), because they were given the opportunity to discuss and give a more detailed explanation of certain aspects of their activities.
- b) In SSH fields, all RUs where the visits of the panel members took place, evaluate the visit positively. This information is usually accompanied by individual recommendations concerning the organisation of such visits, or by

thoughts on the question of a possible logistical solution of a large number of visits, providing the evaluation would take place across the larger number of RUs.

10.3 REFEREES

Answers of referees of excellent outputs to submitted Questions are summarised in the **Background document 10**. Out of the total of 25 referees we received answers from 12 referees working for disciplinary areas Natural Sciences and Engineering and Technology and from 5 referees in Humanities. A concise analysis of the referees' feedback is given in section 7.3.

11. Conclusion

The pilot test of the new method of evaluation of research, development and innovation proposed by the Technopolis Group in collaboration with Technology Centre AS CR, NIFU and InfoScience Prague as a part of the Individual National Project of MEYS “An Effective System of Evaluation and Funding of Research, Development and Innovation” has been successfully completed. The project realisation team of the pilot test performed several particular adjustments to the recommended Methodology that reflected the comments presented in the public consultations, conclusions of the small pilot evaluation, and their own experiences with comparable evaluations and that have also led to adjustments enabling realisation in a limited scope of time.

The adjustments to the Methodology for the purposes of the pilot test concerned mainly the registration conditions of the Research Units, methodological and formal adjustments to the self-assessment report, selecting and contracting the members of the evaluation panels and the referees, certain changes in the bibliometric report, changes in the description of the grades evaluating excellent outputs, eliminating hierarchy of the referees and accompanying academic excellent outputs with bibliometric indicators, and finally, a substantial reduction in compensation of the members of main and subject evaluation panels.

The final recommendations for further adjustments to the future Methodology, specifically in connection with the individual characteristics, are further expanded on at the end of the chapters, always in a separate section under the title “Recommendations”. The possibility of Evaluated Units to, under exceptional circumstances, register more than one Research Unit in one Field of Science, restructuralisation of the self-assessment report, variability of the number of members of the subject evaluation panels and using Czech members alongside foreign experts, the inclusion of also non-scholarly outputs in the criterion “Scientific Research Excellence”, clear formulation and description of the process of calibration exercise and their outcomes aiming to unify the interpretation of the grades, are emphasised as crucial points.

One of the results of the pilot test are the evaluation reports of thirty-one participating Research Units, which were provided to the management of the given Research and Evaluated units, however, were not, as previously agreed, published as they served solely for the purpose of testing the functionality of the evaluation system. At the level of Evaluated Units, the chairs of the main and subject panels collaborated in preparing synthetic reports in the form of a SWOT analysis. The feedback of the members of the expert evaluation panels and Evaluated Units, which include ideas on establishing a future, nationwide evaluation of Research Organisations in the Czech Republic, are an important output of the pilot test.

In their assessment of the Methodology as a whole, the opinions of the members of the evaluation panels (panellists) on the one hand, and the opinions of the representatives of the evaluated research organisations on the other hand, varied significantly.

The foreign experts of the evaluation panels have, by a majority, recommended the proposed Methodology as a right move in the direction towards implementing evaluation standards that are already commonplace in many European countries. At the same time, they pointed out that a system based on mechanical rating of outputs with points is rare and does not fully reflect the scientific level of quality of a Research Unit.

The representatives of research organisations evaluated the new Methodology less unanimously, mainly due to its complexity and challenges, however, despite certain objections; they evaluated their experience to be mostly helpful.

Although some of the Evaluated Units admitted in the follow-up feedback that they had not given the pilot test as much attention as it would have needed if its results had specific, particularly financial, outcomes. Basically, the general usefulness of evaluation of this kind and using its results for research organisations themselves and its management was not doubted. High costs in terms of time and money, combined with the application of the Methodology proposed, as well as a substantial threat of subjectivity within the evaluation system, were assessed as threatening to the financial stability of research and development by the research organisations. This opinion was accentuated even more in the case of Natural Sciences and Engineering and Technology, which have a fully developed and functioning system of international evaluation of quality of publication outputs. As far as Social Science and Humanities are concerned, the new Methodology is considered definitely useful, though it is emphasised that it does not take into account the specifics of various scientific fields enough. The views also vary depending on the type of research organisation. Whereas the university faculties evaluated were mostly critical to the new Methodology, positive opinions prevailed in other organisations. This was particularly significant in the case of institutes of Academy of Sciences CR, which have already had many years of experience with a “peer-review” form of evaluation.

Overall, the pilot test has definitely fulfilled its objective and proved that the recommendations of the Technopolis Group and its subcontractors, obtained as a part of the IPN Project “An Effective System of Evaluation and Funding of Research, Development and Innovation” are useful as a basis for future evaluations, however, with certain necessary modifications (see Implementation recommendations of the project team IPN Methodology, where a number of conclusions taken from the pilot test are also used). After certain changes and utilising the experience with comparable systems of evaluation, such as “Methodology 2013+”, as well as the currently undergoing evaluation at AS CR, the IPN Methodology team recommends implementing the new system of evaluation as soon as possible, in the first trial phase in only a chosen segment of research organisations, i.e. public universities, with a certain though not crucial impact on the level of their institutional funding. This interim period seems to be necessary, mainly due to the complexity, as well as the cost and time needed for the implementation. After successful completion of this phase, a nationwide system of evaluation and funding of all research organisations in the Czech Republic can be established. The KA4 team is, however, aware of the fact that practical implementation of the new Methodology nationwide is, above all, primarily a political decision.

12. List of background documents

Background document 1	Submission Guidelines for Evaluated Research Organisations
Background document 2	Template of Self-assessment Report
Background document 3	Minutes of Calibration Meetings
Background document 4	Adjustments to the Bibliometric Report
Background document 5	Example of a Bibliometric Report
Background document 6	Expert Panel Member Guidelines
Background document 7	Overview Tables of EvU and their RUs
Background document 8	Main and Subject Panel Members' Feedback on Methodology and Pilot Test
Background document 9	Evaluated Units' and Research Units' Feedback on Methodology and Pilot Test
Background document 10	Questions and Answers of the Referees of Excellent Outputs

13. List of terms and abbreviations

term	abbreviation
Czech Technical University	CTU
Disciplinary Area	DA
Engineering and Technology	ET
Evaluated Unit	EvU
Evaluation Management Board	EMB
Evaluation Management Team	EMT
Evaluation Methodology	EM
Evaluation Report	ER
Excellent Output	EO
Expert Panel Member Guidelines	EPMGL
Field of Science	FoS
Full Time Equivalent	FTE
Humanities	HU
Industry & Business Services Research Organisations	IBRO
Information System	IS
Information Technology	IT
R&D&I evaluation methodology based on recommendations of Technopolis Group and its subcontractors	Methodology
National Resources	NatRes
Natural Sciences	NS
Remote Evaluation	RE
Research and Technology Organisations	RTO
Research Organization	RO
Research Unit	RU
Scientific Research Organizations	ScRO
Self-assessment Report	SAR
Social Sciences and Humanities	SSH
Technopolis Group and its subcontractors	TP

Pilot test of new evaluation methodology of research organisations
Final Report

This document has been prepared as a part of the Individual National Project
“Effective System of Research Financing, Development and Innovation“
MŠMT ČR, Karmelitská 7, 118 12 Praha 1
www.metodika.reformy-msmt.cz