Progress report on the IDEA meeting of the Working Group on the

Inclusion of animal testing alternatives into QRA for skin sensitisation

April 26th, 2016

IFRA offices
6, Av. des Arts,
1210 Brussels (Belgium)

1. Introduction

The main purposes of the meeting were to:

- Gather together experts who are active in the area
- Identify the nature of the tests that are already available and/or at an advanced stage of development and their current areas of application
- Discuss how they should be best applied and any general issues in their application
- Discuss how the findings of such studies should be integrated with other data for hazard/risk assessment purposes
- Consider next steps.

The meeting did not discuss toxicokinetic aspects of the induction process since this has been a topic of previous workshops on prehaptens and prohaptens (see summary figure in appendix).
Presentations were given by JRC, Cosmetics Europe, EPAA, David Basketter, Andreas Natsch, CAAT, RIFM, and ECHA.

2. Basis for the development of alternative tests

Both the QRA1 and the QRA2 for skin sensitisation largely rely on an animal test data (mainly the LLNA), in order to identify the potential for induction by a fragrance material. The ban by the EU on animal testing for cosmetic products/ingredients has provided the primary stimulus to the development of non-animal alternative tests to assess the induction potential for induction/sensitisation. It is has been, for the past few years, and is currently, a very active area of research not only for cosmetic ingredients but for other areas of chemical use too.
For a number of years, the LLNA test, which uses lymphocyte proliferation in mice as the endpoint measurement, has been the method of choice. The LLNA was the first fully validated test for the assessment of the induction

potential of chemicals for skin sensitisation. It assesses both the identification of the hazard and the potency. In the USA, for fragrances, the findings are backed usually by HRIPT investigations (see QRA1 and QRA2).
The development of the LLNA was based on an understanding of the process of induction. It is appropriate that the animal testing alternatives should be based on an increasing understanding of the sequence of critical events, starting with exposure, that result in induction. This is in line with the work going on in other areas of toxicology that are increasingly focused on the identification of the key events in adverse outcome pathways/ mechanisms of action (see TOX21 and subsequent research).

2.1. Methodological requirements

The following were noted as important issues for the development and validation of suitable non-animal tests to assess the potential for sensitisation in human skin of fragrance materials:

- Identification of an induction /sensitisation effects assessment data set that is appropriate to use as the gold standard.
- Access to well characterised reference chemicals (training set(s)) for each chemical domain of relevance. These need to be selected on the basis of gold standard data. It is recognised that this requirement may, at present, exclude any representatives of some chemical domains (based on physicochemical properties).
- Knowledge of the adverse outcome pathway(s) /sequence of events responsible for induction/ sensitisation in the skin and in particular the critical events for induction/sensitisation in human skin.
- Availability of biological preparations (tissues, cells etc) which are relevant, well established, can be made widely available, and are stable.
- Relevance of the exposure conditions. This includes suitability of the test system for the addition and retention of the reference chemical(s) and the test chemical(s). Kinetics is one aspect of this.
- The need for a widely accepted framework for the utilisation of all the relevant data generated including characterisation of uncertainties.

2.2. Gold standards

It was agreed that ideally the gold standard is human data, however the data is of variable quality and, if taken from clinical patch testing, identifies elicitation rather than induction. As RIFM noted, HRIPT is also an important source of human data including a defined sensitization phase, however it generally only provides information on a single dose and therefore does not provide dose response data. Furthermore, the data base is not nearly as extensive as that for the LLNA in mice. Consequently, the practical choice as the gold standard for most purposes is findings in the LLNA. An important issue is to check the variability within and between laboratories in the reported findings for individual chemicals.

2.3. Reference chemicals

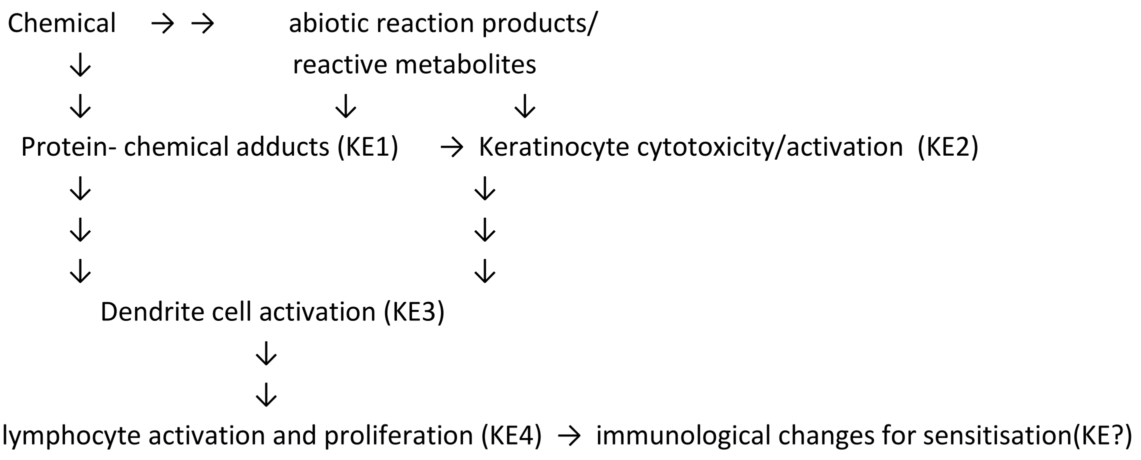
Two aspects were touched on:

- The development and validation of each test system for application to each important chemical domain (based on physicochemical properties). Training set selection should include difficult chemicals although it is important to define what is meant by difficult. Factors that should be considered include physicochemical properties (such as high lipophilicity, chemical stability, ionisation, volatility), and potential for metabolism to reactive intermediates.

- The selection of positive and negative control reference chemicals, appropriate for the chemical domain of the chemical being assessed for hazard/risk assessment purposes.

2.4. Adverse outcome pathways/sequence of events.

There was a general acceptance that test development and acceptance should be based on current understanding of the adverse outcome pathway(s)(AOP) / mode of action (MOA)/ sequence of key events (KE's). The diagram below identifies the understanding of this sequence of events.



Chemical → → abiotic reaction products/ reactive metabolites

Protein- chemical adducts (KE1) → Keratinocyte cytotoxicity/activation (KE2)

Dendrite cell activation (KE3)

T lymphocyte activation and proliferation (KE4) → immunological changes for sensitisation(KE?)

Key event 4 (KE4) is the nearest equivalent endpoint to that measured in the LLNA test. However, test development has concentrated on the preceding critical events KE1, KE2 and KE3.

What is unclear at present is whether this represents the sole mechanism by which induction can occur and/or there are important variants after KE1. It is possible that chemicals in certain chemical classes/domains might act in a slightly or substantially different way. It was noted that most non sensitisers in the databases used for validation of in vitro assays are not cytotoxic, but this may not be true for the chemical universe (rather a database bias).

3. <u>Test systems</u>

3.1. Relevance of exposure conditions

The following aspects were touched on but there was insufficient time to address them in any detail:

- Should the fragrance material alone be assessed or should it include other common components in cosmetic product formulations? An argument in favour of the addition of other components is that they may be responsible for an irritation reaction that promotes the induction by the hapten.
- Under what circumstances is bioavailability a problem? Some participants referred to results from analysis of data showing that bioavailability appeared not to be a concern. This, from a mechanistic point of view, is a rather unexpected finding.

- Is the metabolic capability of the various test systems adequate to elicit a positive response from prohaptens? Again evidence was cited that the metabolic capability of the various test systems was a minor issue since there are very few exclusive prohaptens (i.e. most are either direct acting haptens or are also prehaptens). This is perhaps surprising as the findings for studies on the genotoxicity of chemicals (which also requires that the chemical or a product of the chemical is chemically reactive) indicate that an active drug metabolising enzyme system is often a requisite.

Other aspects that were mentioned but not discussed at all because of time constraints were how to deal with:

- Volatility, highly hydrophobic chemicals, ionised chemicals, generation of conditions for prehapten to hapten conversion, and mixtures.

3.2. Available test systems and endpoints

As an introduction, D. Basketter presented a brief overview of current understanding of hazard identification of skin sensitisers utilising alternative approaches to animal testing. There was insufficient time to discuss the details of individual tests and whether they can only provide yes or no results or can also provide reliable information on potency.

i. Most widely studied

Several speakers identified the following tests and the most advanced and potentially utilisable for hazard identification purposes.

- In silico (e.g. TIMES, DEREK, OECD TOOL BOX, TOPKAT)
- In chemico (Key event 1) eg DPRA (validated and OECD adopted); PPRA, AREc32 under evaluation
- In vitro (Key event 2) eg KeratinoSens (validated and OECD adopted); LuSens (validated peer-reviewed by EURL ECVAM) and epidermis IL18 under validation
- In vitro (Key event 3) eg h-CLAT (validated and OECD adopted); Il-8-Luc, U-SENS, (both validated and peer reviewed)
- In vitro (key event 4) much more difficult to reflect in a non-animal system and not close to a validated system yet.

JRC reported that DPRA, being *an in chemico* assay, has no metabolic capacity. Based on the examination of five known prohaptens, only one of them was detected by the DPRA whereas h-CLAT predicted all of them correctly as sensitisers. It is not known what metabolic pathways are involved. It was noted that if a chemical was also a prehapten the metabolic capability of the test system is not a limiting factor for hapten formation.

ii. Other test systems

Many other test systems are in various stages of development or evaluation including: EE potency assay, IL-18, SenCeeTox, VitoSens, GARD, SensiDerm, mMUSST, PBMDC. 3D keratinocytes based models with a differentiated epidermis and stratum corneum are of particular interest because doses can be applied topically and may provide more reliable dose response data. SENS-IS, which uses Episkin as the test system is an interesting development in that it relies on the expression of a large panel (~200) of genes by chemicals as the endpoint –approximately 90% prediction accuracy has been claimed. The method has been validated but has not yet undergone independent peer-

review. A claimed benefit is that it can apply to hydrophobic fragrance materials. The identity of the genes assessed is difficult to identify unfortunately, because of patent /commercial issues.
Many of the tests under development or even commercially used appear to cover the same key events. In this respect, EPAA informed about a current collaborative project to test ‘difficult substances’ in three different 3D tissue models.

iii. Test selection

With so much commercial interest a particular challenge is to ensure an objective method for the evaluation of individual tests (which is well managed by ECVAM) but also to identify for each test:

- The chemical domains for which it is suitable
- Other constraints in terms of the conduct of the test
- The added value of the test over other tests for the same critical event endpoint measurements. Some evidence was presented that indicated that KE3 based tests may not necessarily provide added value to the findings from KE2 based tests. It is important to establish whether this is a generally observed conclusion.
- The incidence of false negatives and false positives. It was noted that if a tiered approach is employed for testing purposes, false positives are much more acceptable than false negatives in the first tier. However, the comment was made that neither false positives nor false negatives were often observed.

4. Prediction capability of the current tests

4.1. Hazard identification

The JRC presented the data below from Urbisch et al. 2015 which indicates the range of performance for yes/no findings for four tests DPRA, KeratinoSens, LuSens and U-Sens as shown, using 150 chemicals from 5 different LLNA potency classes. The table presented by EURL ECVAM data provides good evidence that most prehaptens are identified by the in vitro tests.

Parameter	In vitro vs LLNA	In vitro vs human	LLNA vs human
Sensitivity (%)	74-86	72-89	91
Specificity (%)	68-85	64-84	64
Accuracy (%)	74-81	82-84	82

No bench mark was proposed at the workshop as to what should be deemed as an appropriate value for acceptability for sensitivity, specificity and accuracy. However, the above findings compare human data with LLNA data. Since LLNA findings are widely accepted for human hazard assessment purposes and the aim of the in vitro test development is to replace animal tests, in particular the LLNA, the comparison of this test with human data provides a very relevant practical benchmark for acceptability. In comparing in vitro tests with either LLNA or human data it should be noted that the in vitro tests that were employed in these studies are based on one specific mode of action (OECD ENV/JM/MONO(2012)10) leading to induction.
Cosmetics Europe summarised collaborative study findings for some of the individual tests in comparison with data using the LLNA and human data and based on almost 100 cosmetic ingredients belonging to six human potency

classes. The accuracy for single methods compared to LLNA ranges between 68,4 – 78,6% and when compared to human data 73,6-76,5%.

The findings indicate that whether the comparison is with the LLNA data or human data the three parameters of sensitivity, specificity and accuracy are rather similar, thus justifying the use of the LLNA findings as the gold standard for practical purposes.

4.2. Hazard potency

A fully quantitative assessment of potency is essential to replace the LLNA test in the QRA. The available in vitro tests were agreed to be satisfactory for the identification of the hazard of induction from chemicals. However, none are yet suitable to properly identify potency even in a qualitative manner. This is true too for in vitro tests that have been developed to mirror other in vivo adverse effects. There was a discussion on how sophisticated the potency estimation needed to be for CLP purposes (see below for likely categories), according to 2nd ATP to CLP (Regulation EU No. 286/2011). The most promising way forward using the currently available tests appears to be:

- to critically evaluate the chemical reactivity,
- to compare test findings for each chemical of interest against the findings for selected reference chemicals,
- to concentrate on those tests that provide a dose response data and are complementary to one another.

5. Use of other relevant data and weighting of findings

i. Chemical reactivity

Understanding the chemical reactivity of each chemical of interest is a crucial element of the hazard assessment. The use of QSAR's and characterisation of chemical groupings and read across have a particular role for this aspect of the assessment, next to a quantitative measurement of reactivity with target nucleophiles.

ii. Is exposure of consumers likely to exceed the threshold for induction?

Utilisation of a threshold of toxicological concern (TTC) approach can be used for this purpose (see for example Safford 2008, 2009, 2015, Roberts 2015 in Reg. Tox. Pharm). If a chemical is identified as unlikely to reach the threshold the question arises as to whether any further testing is required.

iii. Biological plausibility

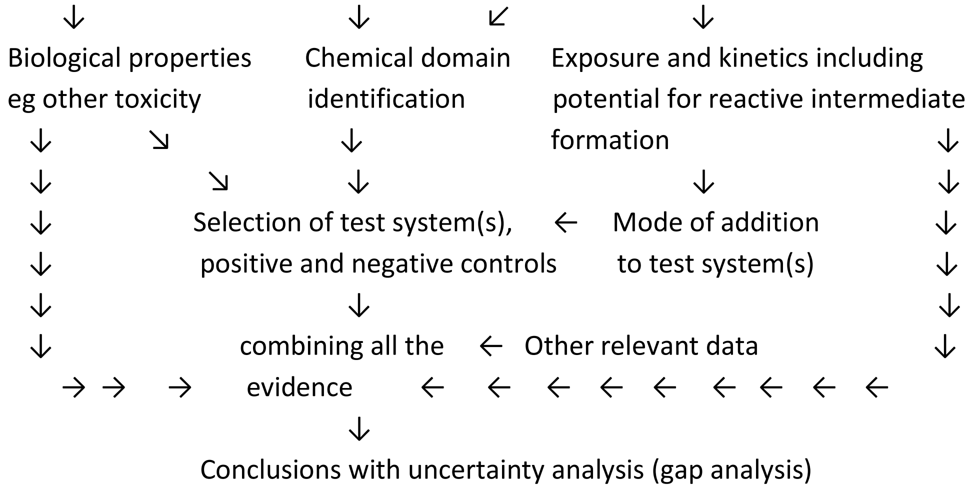
This should be based on current knowledge of AOP's /MoA's together with a comparison with other studies on the same and related chemicals.

6. Integration of all the findings

The following figure summarises the required approach.

Data mining ↔ Chemical and its use(s) ↔ Physicochemical properties

↓ ↓ ↙ ↓



Biological properties eg other toxicity

Chemical domain identification

Exposure and kinetics including potential for reactive intermediate formation

Selection of test system(s), positive and negative controls ← Mode of addition to test system(s)

combining all the evidence ← Other relevant data

Conclusions with uncertainty analysis (gap analysis)

6.1. The challenges for an integrated approach

As noted above there are various potential types of data that should be considered. As discussed, there are also various ways in which findings can be combined. Whilst JRC summarized the defined approaches for potency prediction being considered within the context of the OECD IATA GD, an alternative approach was mentioned by CAAT. Further, A. Natsch described in greater detail some early attempts to reach quantitative potency using a combination of parameters from in vitro tests. Unless there is a widely agreed, formalised procedure for evaluating the findings, each integrated approach involves an element of expert judgement but is also largely transparent. Namely:

i. Algorithms/decision trees. The benefits of such a system are simplicity and its suitability for the use of a tiered approach to testing and analysis of findings.
ii. Machine learning techniques such as Bayesian or neural networks. Both have the advantage of enabling linkage to be identified between a substantial number of findings, showing situations where findings reinforce one another and where this is not the case.
iii. Statistical techniques such as regression models. These have been used to predict points of departure for quantitative risk assessments.
iv. Weighting of findings from particular tests based on their relevance and reliability (weight of evidence methodology). This has the advantage that it assesses both the quality and relevance of test findings and that it is increasingly being accepted by regulatory authorities as the methodology for hazard and risk assessments.

Utilising machine learning based on one or more of these assessment schemes is likely to enhance their value for hazard/risk assessment purposes. The challenge is to identify the most appropriate methodology in order to estimate potency.

In situations where all the data indicate that a chemical is an inducer/sensitiser and there is consensus on potency, any of these methods can be easily applied to reach conclusions for regulatory or other purposes, although some estimate of the degree of confidence may be helpful.

The challenges arise when the data appears to be inconsistent. This is likely to require weighting of individual test results based on reliability and relevance. There was insufficient time to discuss this important issue.

A further important issue that was only briefly touched on, was how to characterise and express uncertainty in the analysis.

7. Utilisation of the findings for regulatory purposes

7.1. Hazard ranking

Hazard identification is, as discussed above, based on findings in tests that have been developed through an understanding of key events in the induction/sensitisation process. Most of these tests are to provide a yes or no response and the initial intention in their development was that they would be used for hazard assessment purposes in conjunction with other tests probably in a tiered approach.

It was agreed that it is important to identify redundancy between tests. It was stated that combining findings on reactivity and KE2, or reactivity and KE3, gave similar correlations while combining all three gives little improvements. If this finding is typical it indicates that under the conditions of the test once covalent binding to the target protein(s) has occurred, activation of dendritic cells is almost inevitable. This implies a very high level of redundancy in tests for KE2 and KE3 and places particular emphasis on KE1 tests and the importance of developing KE4 tests.

For classification and labelling purposes a categorisation scheme is necessary based on potency. The ECHA representative emphasized that a reliable way to address potency is needed as soon as possible. Therefore, the primary need of the integrated testing methodology is to classify each chemical. The most likely is a CLP scheme such as:

Extreme- strong 1a
Moderate-weak 1b
Weak/very weak-
Non sensitiser. NC

While the in vitro test appear to be able to distinguish between non-sensitisers and moderate/strong sensitisers (assuming there is a single MoA) it appears at present to be problematic to differentiate between moderate and weak sensitisers. It is unclear whether the tests reliably separate weak and non-sensitisers. Until these issues are resolved it must be concluded that the currently available methods are not sufficiently developed to be used for categorisation purposes based on potency. This problem also applies to some extent to the LLNA test.

7.2. Risk assessment

How to use animal testing alternative methods for risk assessment, in the absence of any animal or human data, remains a challenge. Until quantitative potency can be confidently established at exposure levels that are relevant to that of humans the alternative methods will not be utilisable for risk assessment purposes. An interesting approach, which was not discussed at the workshop, is the study by Maxwell et al (2014, 2016). However, this describes a study in which a number of volunteers were tested with the same chemical and is therefore unlikely to have wide applicability.

8. Conclusions

The workshop was very useful in characterising the current state of the science. It demonstrated very clearly that a great deal of work is being carried out on in vitro alternatives for identifying skin contact allergens. Since it was an IDEA initiated workshop the focus was on alternatives for fragrance ingredient testing as a key component of the QRA, although work on the testing of other cosmetics ingredients was also referred to. It would be very helpful to compare notes on progress with scientists working on alternative development for contact allergy from other industries. The general conclusion from the workshop was that hazard identification can be achieved reliably in the near future using alternatives to animal testing. However, assessment of potency is still under development and to date there have been only limited attempts to achieve this. Consequently, the available methodology is not sufficiently developed to enable the LLNA test to be replaced for QRA purposes.

9. What is needed

A number of issues were raised that need to be addressed to enhance the development of a non-animal testing QRA. These include:

i. Agree and use consistently a common terminology.

ii. Agree and ensure that the most appropriate machine learning technique are utilised for data assessment purposes.

iii. Develop further in chemico tests to improve further models and tests for chemical reactivity prediction.

iv. Extend the HRIPT data base as the gold standard for human sensitization.

v. Identify and ensure the availability of key benchmark chemicals, (reference point LLNA and/or human data) and utilise them for studies on chemicals in the same domain in suitable tests.

vi. Agree a generally acceptable framework for incorporating all the data in order to reach a conclusion which is likely to involve the ranking of tests related to their chemical domain and use. It will need to embody means of characterising uncertainties.

vii. Identify, and give emphasis to, existing and potential tests that are most likely to provide the best opportunity for estimating potency.

viii. Ensure awareness of new developments in the understanding of AOP for induction with particular attention to alternative modes of action to those on which the current tests are based.

ix. Provide worked examples for chemicals in different domains comparing LLNA based and in vitro based hazard characterization.

x. Identify and characterise uncertainties/variability.

xi. Address the issue of how to deal with difficult chemicals and mixtures.

10. Potential roles for IDEA in progressing the development and use of alternatives

- Serve as organiser/co-organiser of meetings to review findings, identify needs and prioritise them.
- Ensure stakeholders (in particular the Commission and the SCCS) are regularly updated on the progress.
- Act as a reference point for who is doing what in the industry, research organisations and collaborative partnerships.

- Provider of reference chemicals.
- Source of reference data (LLNA and human).
- Support (financial and /or administrative) for specific research and/or development topics.

11. Proposed action

A small expert working group should be established by IDEA to prioritise the way forward and the role of IDEA in achieving it.

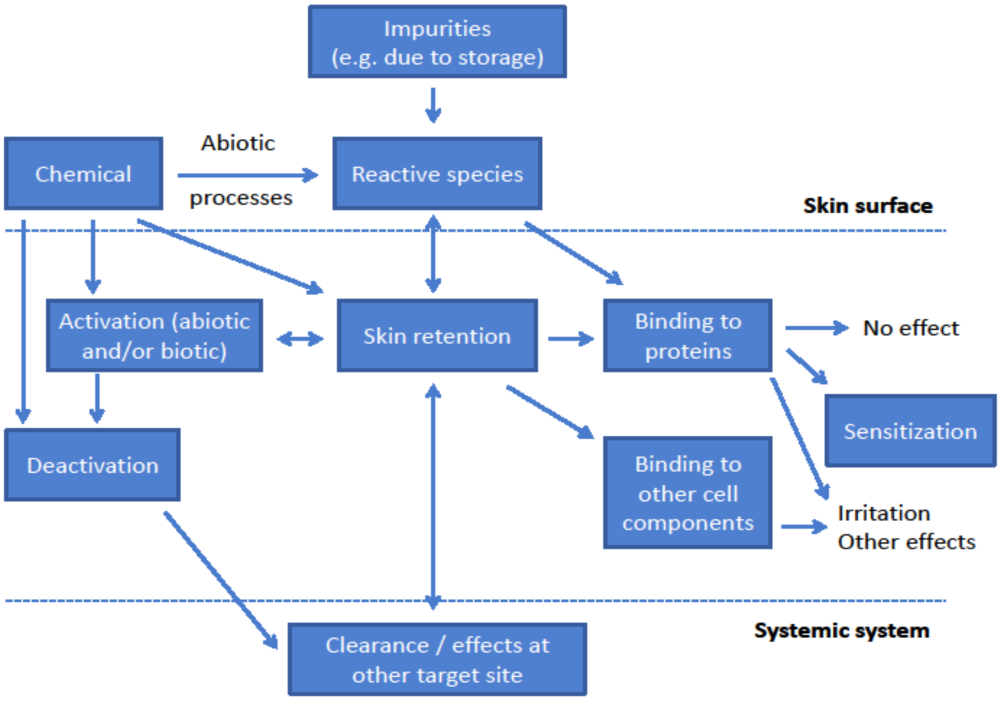


Figure 1: Distribution, transformation and effects of pre- and pro-haptens.

Appendix 1: Participants to the meeting

Anne Marie Api (RIFM)
Karin Aschberger (JRC)
David Basketter (Consultant)
Silvia Casati (JRC)
Bruno Hubesch (CEFIC LRI)
Martina Klaric (Cosmetics Europe)
Irene Manou (EPAA)
Andreas Natsch (Givaudan)
Laura Rossi (ECHA)
Costanza Rovida (CAAT Europe)
Izabela Taborska (EU Commission, DG Grow)

Pieter-Jan Coenraads (Observer on behalf of SCCS)

Amaia Irizar (IDEA Management Team)
Matthias Vey (IDEA Management Team)
Cécile Gonzalez (IDEA Management Team)

Hans Bender (Moderator of the IDEA Workshops)

Jim Bridges (IDEA Supervisory Group)