International Initiatives to Develop a Global Strategy On Endocrine Disrupters Testing and Assessment

Herman B.W.M. Koëter

OECD

For those of you who are not well informed about OECD, it is the Organization for Economic Cooperation and Development. We have thirty members at this moment: North America, all European countries, Asian countries, and the South Pacific. One may wonder why OECD would be so interested in endocrine disrupters work, and in testing in general, since it is an economic organization. The basic reason for our involvement in testing and assessment is that OECD could contribute to the harmonization of procedures as applied in our Member countries: procedures for registration of chemicals, procedures for notification, but also for risk management. Obviously in many areas but most notably for industry, for those who produce the chemicals, harmonization will make it easier to provide data packages that would be acceptable in all countries, that would facilitate import/export and, consequently, would contribute to substantial economic savings.

We basically build our work upon two different objectives. One is to protect man and the environment. This is the cornerstone of the work. We try to do that with the highest quality instruments possible and that is the subject we are talking about this afternoon. The second objective is to increase efficiency and with that, economic growth in our Member countries as well as in non-member countries.

Elements that contribute to that efficiency are methods of sharing the burden so that there is neither duplication nor unnecessary work done by countries that has already been done elsewhere. Furthermore, we would also like to avoid non-tariff trade barriers.

Initially, I had a few slides that I am now going to skip, considering the time, but because of being chairman of the session, I just told all of the speakers that I want them to restrict themselves to the allotted 20 minutes. The first thing that I should then do is to stick to the 20 minutes myself.

What I do want to share with you is a few statements of European Committees, and in Europe, the European Commission makes use of a number of Science Committees for various areas. Two of those have had several meetings on endocrine disrupters, and the means in Europe by which the Commission should deal with this issue. One of these committees is the very broad scientific committee on the toxicity and ecotoxicity of the environment.

In 1999 they stated in one of their most recent reports that at present the Test Guidelines cannot detect the endocrine disrupters effect sufficiently. They added that tests need to be announced and new tests need to be developed.

If I look at another committee, which is the scientific committee on plants that basically deals with pesticides and everything that comes with crop protection, they stated that ecotoxicological risk arising from endocrine disrupters can be captured generally by the current assessment scheme. In other words, we have two scientific committees in one part of the world that have very different, though not really opposing, views on defining the problem and its size.

In 1997, IFCS, the Inter-Governmental Forum on Chemical Safety which comprises about 120 countries in the United Nations that come together on a regular basis, recommended that these five series of activities needed to be conducted.

Definitions are needed and they need to be harmonized, as well as the terms of use in the area of endocrine disrupters.

The research should be coordinated.

Gaps and testing priorities should be identified.

Test Guidelines should be harmonized for that matter.

The inventory already in place in the US should be broadened and made more worldwide and kept and maintained, and information should be exchanged.

They requested that these five issues be dealt with by IOMC, the Inter Organizational Program for the Sound Management of Chemicals. This is a mouthful that essentially means that a number of UN organizations, that in one way or another deal with chemicals, like WHO, UNEP, UNIDO, UNITAR, ILO and OECD, work together as organizations to carry out the work that their countries had requested be done.

Looking at the five priorities, the first one has been taken on jointly by OECD and IPCS: it is a joint project on definitions of terms. The co-coordinated research is a project that has been conducted by IPCS. Later in this meeting Bob Kavlock will go into detail of the outcome of that particular work, identifying the testing priorities, the gaps and the Test Guidelines that were officially an OECD task.

An inventory of their research activities was again left to IPCS, who has now handed it over to the commission in ISPRA, Italy, where the database is maintained. Information exchange is done by each of the participating organizations and it is done in a rather structured way in OECD. So these are then the organizations in IOMC that work together to jointly agree on the work.

So how did we pick it up in OECD? First, we established a task force — which is usually referred to as the EDTA, the task for Endocrine Disrupters Testing and Assessment — with a number of objectives.

Again, the objectives are rather obvious: information needs to be provided to coordinate activities, member countries' work needs to be harmonized, and that information should be available for other countries as well. The next objective of revising Guidelines that already exist and developing new Guidelines is, of course, an important issue. The last objective is harmonizing the hazard parameterization and assessment approaches.

This last objective is much more difficult because countries have very different views on how these chemicals need to be assessed: how much do you need to do and on which types of chemicals should you do that work? Consequently, sharing the burden of testing and assessment is going to be a rather important part of my presentation because I feel that that is the part where we really see the results of all the work.

For you to have an understanding where this is all comprised in OECD, take for example the Test Guidelines Program that I am directing; it is an enormous program in OECD containing numerous expert groups. There are between 6,000 and 8,000 registered experts that are nominated by member countries as being experts that have to work in OECD, and in addition to these thousands there are many more who, on a voluntary basis, also provide their comments and suggestions. They are all aligned under what we call the working group of national coordinators of the Test Guidelines Program, and under that there is the EDTA with two subgroups that I will talk in a minute.

When the EDTA started, their task was to enhance and identify those Guidelines in OECD that could probably be used also in the detection of endocrine disrupters, and to see if those needed to be enhanced, improved, or updated to make them better suitable for this task. But also, new Test Guidelines, new methods, and new screens probably needed to be developed, and if the development was done, they need to be validated, which I will address later.

The strategy that one needs to develop in order to apply all of these tools that are available was an important part, as well as the sharing of the burden. Obviously, the work in OECD countries is not directed by the Secretariat in Paris; my staff and I carry out the work that countries ask us to do. Countries do not always ask that via their prime minister of course; they make the requests through their agencies that are involved through the work and through the experts in their agencies who represent their country in the various groups. The work is based on consensus, which is simultaneously an advantage and a

disadvantage. It is obvious that the advantage is that if you have reached consensus and everybody is happy with it, that you have a broad, global acceptance. The disadvantage is that scientists tend to disagree and that it sometimes takes years, if not longer, before we finally reach some sort of consensus.

The next slide is an important slide and it is actually one of two slides in which we try to bring together what we call a conceptual framework for the testing. You can imagine that as soon as this problem of endocrine disrupters came up, experts and NGOs immediately started to say, "You need now to test all the chemicals extensively and definitively so that we have a real answer," and so on. Others were saying that that is not a realistic approach.

We tried to develop a testing paradigm that everybody could accept. This is a modified version of an earlier one. And although we really have not discussed it in anymore detail, this is basically an agreeable concept in which you will always start with collecting the information that is probably available on a number of chemicals and see if you can come to conclusions based on that information. If not, then you have to sort all of your chemicals, meaning about 80,000 chemicals all together, which is a lot. They need to be sorted before you can start decent testing, and sorting simple tests, i.e., high throughput screening (and we will have sessions about that this afternoon as well) and *in vitro* tests, are considered as popular tests for that particular early stage of work.

After that you may well consider, for certain chemicals, moving straight into what is called longer-term or definitive tests. You can also pick some short-term tests first to see if this is really a problem, or if it can be moved a little bit down on the list of priorities. Of course, we try to develop both at the mammalian and ecotoxicology side tests in the area of short-term screens. We do the same for long-term testing, also in mammalian and ecotoxicity. We have identified a series of tests, and I will run through some of them in a little bit more detail. This is to show you that basically all countries have bought into this scheme. Some countries prefer to go through all the chemicals in one way and other countries in another.

The United States is one of those countries that says we should go through all 80,000 chemicals. Meanwhile other countries start at the other end of the scale and prefer to start with those chemicals that, for one reason or another, have attracted attention. Obviously, we will end up doing a little bit of both and see how we can progress.

The new tests that have been identified so far are the ones that are listed here: the Uterotrophic assay, which is a very simple assay for estrogen effects; agonists and antagonists should be picked up by that. The Hershberger assay is a similar assay but for androgens. Then we have the fish screening test, the fish full life cycle test, an avian two-generation test, and amphibian screening and testing. These are all recognized as new tests; we have no available Guidelines for them.

One question might be, "How were they selected"? That was in itself a major process done by countries like the United States and Japan who have been very actively involved with various expert groups in their countries to come up with a selection of tests that seemed to be promising in that area. Others had to go through the scrutiny of detailed background review papers of the literature to see which tests were promising. So far these examples have made it; there are more to come in the future.

The Rodent Uterotrophic assay: I will say a few words only on this test and will not go into detail. Dr. Owens, the next speaker is, first of all, much more knowledgeable than I am on the subject, and it is part of his presentation anyway. The Uterotrophic assay is a test in use since the 1930s; it evaluates the ability of a chemical to mimic estrogen activity in an intact animal, which is important.

The status of this assay is that the validation work is completed; I say that in one simple sentence, but it took about three to four years altogether from start to finish. I will not give you the dollar figure that was involved for that, but will say that it was really substantial. Twenty-five laboratories all over the world carried out a number of those tests to make sure that it worked.

The Hershberger, as I said, is a similar assay of a similar level of scrutiny. It looks at the effect of chemicals on the restoration the rate of the prostate, as a gland, and the seminal vesicles in castrated animals, to show the effect of androgens. The validation work is in Phase II. I will talk a little about validation in a minute. Obviously, Phase II comes after Phase I, but Phase II is not a really well-defined phase and neither is Phase I; it all depends on what you find in your earlier assessment while you move on in your validation.

Then we have the 14 or the 21 day fish assay. We actually haven't yet really started with that; it is only recently that we picked this up. It is based on an existing OECD Guideline 204 and 215 that uses adults and juvenile animals. We have proposals for these assays from the United States and from European experts, respectively, and we are now trying to find out which works best, 14 or 21 days; also it depends, of course, on the end points that you measured. We have to do some experimental work to find that out. It should also be applicable to different fish species. Obviously a country like Japan has a preference for fish that are indigenous species here. Australia, North America and Europe have other preferences. So there are at least a few fish species involved in the work.

It is supposed to be sensitive enough to pick up both the estrogen and androgen activities and the validation work is about to start. The new elements in these tests are vitellogenin, as the most important marker, and even that needs testing in itself to come up with the reliable test kit. We are also focusing on gross morphology in fish and gonad histology that have not really been conducted before.

The fish full life cycle assay is obviously a test that finds its niche somewhere down in the testing paradigm. It is an expensive test. We currently have tests under way that are being tested by Germany, Japan, and the United States, using two to three months' exposure. Again, these are fish that we use also in the short-term tests; we use the same elements, but it is a longer test and it exceeds one generation. It thus includes reproduction, so it is obviously a more definitive test.

Avian reproduction tests: we have two particular assays here. We have a two-generation test, which is a new one; there are no Guidelines. Provisional Guideline 206 is an existing Guideline for one generation in Japanese quail and bobwhite: I will not go into the details, but these birds are in many aspects very different, in such a way that you have to design your protocol according to their hatching time, etc. There are all kinds of differences between these two; you cannot just use the same protocol, and it does not matter whether you pick the Japanese quail or the bobwhite.

Fertility parameters have never really been covered in the two-generation assay, at least not in a sophisticated way; the same holds true for reproductive toxicity endpoints. I have to say that the expert meetings that we have had on birds and on fish date back several years already, and involve meetings in Japan, in the United Kingdom, the United States, SETAC, and so on.

For some reason that environmental toxicologists are unable to explain to me, environmental toxicologists tend to discuss and discuss and discuss, whereas mammalian toxicologists, in my experience, discuss and then come up with solutions. So in one way or another, they are more pragmatic than environmental toxicologists.

Frog metamorphosis assays, lastly, are still very much in their infancy. The frog was flagged as an important species to be considered in the whole environmental assessment of endocrine disrupters. The *Xenopus* is obvious as the species of choice. There is already much work done on that, but still we are in the midst of scientific discussions on this assay.

If I move onto the mammalian work, we have Guideline 407: it is a four-week systemic toxicity test that is being enhanced, and I have a slide or two on that. Guideline 416 is a two-generation reproductive study, and the avian study I already mentioned.

The 416 two-generation study was revised, updated and adopted again in January this year and we are starting again with that Guideline. It was updated based on parameters that were already considered

five or ten years ago that need to be included, not necessarily all focusing on reproduction and fertility, and we are now looking again at that Guideline with a view to fully covering endocrine disrupters.

Guideline 407 was selected not because people found that this is an absolutely great Guideline, but because it was seen as a matrix where you have systemic exposure and where you have not too big a number of animals in each group. Scientists came up with the thought of enhancing that Guideline in such a way that you would include reproduction of fertility parameters without mating of the animals.

If it would be possible to use other successfully sensitive means the result would be a good assay that is useful in large parts of the world. It would be used in early assessment as a preliminary indicator of whether there will be an endocrine disrupter problem or not, without having to do all kinds of other tests. The validation work of this assay has been led by industry worldwide. The test optimization phase was the first part of the validation work with two protocols. Protocol A was actually the current 407 plus all of the additions that I will mention in a minute. Protocol B included only the essential elements of 407. The essential elements basically meant that all of the functional testing and the detailed histopathology of organs other than those dealing with reproduction or in any way with the endocrine system, were not used in that protocol in order to make the exercise a bit cheaper. If you do the addition it is really an expensive assay.

The endpoints that were added to Phase I of the validation study were the weight of the pituitary, the thyroid, the ovaries, the uterus, the prostate, the individual testis, the epididymis, the seminal vesicle, and the coagulating glands. The histopathology of these same organs and a few more was also conducted.

The number of hormones — we heard this morning about the thyroid hormones — these were all added to that assay, and it had never been done before. Spermatology was included: the number of sperm cells, the motility, and the morphology were all counted, and the estrous cycle was assessed.

After the first phase was done and the tests were evaluated, a couple of chemicals were identified as obvious endocrine disrupters. Both estrogen and androgen were tested in the first phase (I did not bring any slides on the results because I did not think it was relevant for my presentation). The outcome was that in Phase II we would include all of the organ weights that we did in the first phase, as well as the histopathology and the estrous cycle.

With respect to the hormones, only the thyroid hormones were left. The others were no longer included. The reason was that hormone determination in a four-week assay was not a sensitive parameter. The variety was too big and the spread was too wide; the effects did not really show up.

For obvious reasons, that is always in hindsight, of course, if you go through the literature you see that in rodents, hormone levels change dramatically due to noise in the room, picking up an animal, temperature, and what have you. All of the above contributed to hormone levels that went up and down over time, and that was not really doable.

On the other hand, at least a few countries felt that we should keep the thyroid hormones in the protocol. They were the most stable of all the hormones, and we did not have any other parameter other than the weight of the thymus. We did not have any other parameter that would be an indicator of thyroid function.

The spermatology was not really a very sensitive endpoint either, but some countries felt that we should keep it in place, but skip the most time consuming, thus the most expensive, end point of morphology because it did not add anything that you had not already learned from the number and motility.

So that is, in a nutshell, where we are, and you are not supposed to ask me any questions about that as the later speakers will explain all of the details. This was simply an overview of what we have done.

I want to focus a minute or two on the validation that I quickly skimmed. The important concepts of validation are, of course, to show the reliability and the relevance of the assay. The reliability is the

relatively easy part because that is where you compare your assay between laboratories. The more laboratories you include, if they all come up with exactly the same test result, the more reliable your assay is.

The relevance is more difficult because it has to tell you the outcome of that test, as well as what it really means with respect to what you want to know. We would like to know how endocrine disrupters affect humans. But on the other hand, we also would like to know how endocrine disrupters influence the environment and the species living in the environment.

Sometimes when we conduct a test in a particular species, the species in itself is the target: the fish, the birds, and the amphibians. But sometimes these same species, certainly the rats and the rodents, are species to be used to extrapolate to the humans. In order to describe for each of those assays in clear terms, what you really think that assay means for humans, and how you can check that in your validation study, is a more difficult part of the work.

There are principles or criteria that are put on paper. This is done by centers in the United States and in Europe, ICCVAM and ECVAM, two more acronyms that I do not want to bother you with. Their principles were brought together in 1996 in Solna, which is a town in Sweden, where the OECD held a meeting to talk specifically about validation and how one should go about it.

It was also considered that validation should be flexible. You can only allow flexibility in that work if it is at the same time transparent. So that was the other element that was considered important. Management committees also needed to be established.

I will not show all those details. We have a validation management group established, one for mammalian tests, with a lead laboratory for the Uterotrophic, the Hershberger, and 407. For the participating laboratories that is more or less the structure. We did the same thing for the ecotoxicological work where we now have two lead laboratories, so we are in the process of assigning the two lead laboratories for the fish screening and for the fish full life cycle test.

Lastly, sharing the burden of testing and assessment: I mentioned already that we have many chemicals that all need to be assessed in one way or another, and probably before we are all dead; that needs to be done in the next five, ten, fifteen years, at the most. The numbers of screens that we have is tremendous, and there is also the time pressure, the costs, the philosophies, and finally the question “how can you bring this all together?” arises.

I will not discuss the chemicals that we have, but if you look at the 80,000 that are in commerce, you could probably subtract about 30,000 of the polymers that most people think are not really chemicals of concern. The number really is small, but still it is tens of thousands of chemicals.

So we can group chemicals; that is one thing. We have the tools that we were talking about and the coordination of the testing; when a country decides that a particular test for a particular chemical is needed, they will share that result with another country so there is less duplication.

The grouping of chemicals is one of those elements that we have now started on in OECD, where we have individual chemicals that have triggered regulatory interest. In Europe, we have such a list. In Japan, there are ministries that have more than one list. There are other ministries that have no list at all, and other places in the world have lists; we bring all these lists together with their criteria on dedicated websites.

Once we have the priority lists for chemicals and their criteria, this will help other countries in deciding whether they have chemicals that they would like to consider. With respect to the tools, High Throughput will be discussed here; we know that Japan has a High Throughput Screen. In the United States another screen has been designed.

Rather than these countries doing their own screening, we now have arranged for them to work together. Japan will screen thousands, and the United States will also screen a considerable number of chemicals. That information will be exchanged so that the whole world can use it.

DNA micro-arrays, QSAR models, and prepubertal assays are the assays that are currently being considered, not only necessarily already in OECD, but by countries that share that information with others.

The coordination of testing means that if there are testing plans, the earlier a country indicates those testing plans, the easier it is for others to say, "OK, if you do this test then we will do that test and together we make up and share the results." This I already mentioned.

Risk assessment is the most difficult part, because countries still have very different opinions about risk assessment. Obviously, that is in part because countries have different social, cultural, and economical situations that contribute to assessment. They also sometimes have different scientific views. So we are not out to try to harmonize all the assessments.

What we try to do is this: if a country does make an assessment, that assessment will be made available to others in such a way that the assessment could be easily reviewed. We are addressing common elements that need to be available in all of the assessments; it will be a structure of an assessment report and will be put up on the web.

Japan is the first country now that has submitted two assessments that are put up on our web pages. In OECD, we have now changed our websites and we do not yet have for each of the pages a user friendly URL. We still have this awful long list here, but that one brings you to the site where you find the assessments that are shared between countries.

I would like to leave it at that for now, maybe one or two questions at the most before I hand the floor to the next speaker.