ECONOMIC DEVELOPMENT AND INFRASTRUCTURE COMMITTEE

Inquiry into Improving Access to Victorian Public Sector Information and Data

Melbourne — 27 November 2008

Members

Mr B. Atkinson
Ms C. Campbell
Mr P. Crisp
Mr D. Davis
Mr B. Tee
Ms M. Thomson
Mr E. Thornley

Chair: Ms C. Campbell
Deputy Chair: Mr D. Davis

Staff

Executive Officer: Dr V. Koops
Research Officer: Ms Y. Simmonds

Witness

Mr D. Groenewegen, Deputy Director, Australian National Data Service.

The CHAIR — David, I welcome you to the Economic Development and Infrastructure Committee's public hearing on the Inquiry into Improving Access to Victorian Public Sector Information and Data. I would invite you to state your name, your address, whether you are appearing in a professional capacity or a private capacity, and if you are appearing in a professional capacity, where you work and your position in that organisation.

Mr GROENEWEGEN — My name is David Groenewegen. My address is 14 Dunloe Avenue, Mont Albert North, and I am here in my professional capacity as a Deputy Director of the Australian National Data Service, which is based at Monash University in Clayton.

The CHAIR — Hansard will be providing the transcript, and evidence taken will become public in due course. Over to you.

Mr GROENEWEGEN — I would just like to give a brief summary of ANDS, what we are and what we are trying to do, because I think that is of relevance to the inquiry, and then to discuss some of the issues that ANDS is trying to address and how they might be of relevance as well. Then I am happy to answer any questions that you might have.

The Australian National Data Service, or ANDS, was established this year only in the last few months, although there has been an establishment project running for most of the year. We are funded by the Federal Government as part of the National Collaborative Research Infrastructure Strategy, which is also known as NCRIS, as one of the Platforms for Collaboration projects. We are still very much in the setting-up phase of the project and putting the staff together. However, we have done a fair bit of planning on what we are trying to do and how we are trying to do it, and so I will talk a bit about that today. We are currently funded until the middle of 2011 — until 30 June 2011.

The aim of ANDS is to try to create what we call an Australian data commons. The idea is that there will be a way that Australian researchers can share and find data that has been produced in their field of expertise, and we will try to provide tools to researchers that will enhance the creation of that data commons and the discoverability of things within that. We are also trying to improve the nature of data management within the research sector across the country, as it has become clear in the last few years, with the large growth in the amount of research data being produced electronically, that data management techniques for looking after that data have not kept up with the ability to produce large amounts of electronic data.

We are also interested in trying to create a culture of data sharing to encourage researchers to share the data they have produced in a format that they are comfortable with. We are not necessarily trying to force all data to be shared in all circumstances. We recognise that there are subtleties and nuances to that, but we feel that everyone's research would be enhanced if more data were made available and shared as publicly as possible within the constraints of the sort of research that is being done.

We are also hoping to increase the capability for understanding how to deal with data. At this stage it is our understanding or our perception that there are a limited number of people with actual skills in the management of data, and we feel there is a need to grow this area, because as more data is produced and more people are producing large amounts of electronic data, there will need to be more people to look after that and to help people to look after that. We do not really want a situation where researchers are spending their time thinking about data management. We want researchers spending their time thinking about research, and we would like data management to become a more straightforward part of that process.

The CHAIR — Can you explain what you mean by that?

Mr GROENEWEGEN — By data management?

The CHAIR — No. Can you explain what you want the researchers to concentrate their time on?

Mr GROENEWEGEN — When email was introduced 10 or 15 years ago, people were uncomfortable with email and how it would work within their normal work practices, and now most people have a way of managing their email — the way they deal with it and how that works in their life — and we think that data management should become that sort of process. You should not really have to think too hard about data management. It should be something you do as part of your everyday life, it should be looked after and someone else should be doing the background stuff necessarily. You would just have to follow some sort of reasonably straightforward process that would allow your data to be managed and made available within that.

The CHAIR — So it is your personal data you are talking about, not general? If you are a researcher and you need information on topic A, that is not what you are talking about; you are talking about when you have generated your research.

Mr GROENEWEGEN — Yes. It is primarily about the research you have generated yourself; that is right. It may well also have to do with data that you have gotten from elsewhere. We think there is actually considerable work to be done to better share data. Part of what we call data sharing is actually doing that more efficiently. We are aware of instances, for example, where let us say a lot of people use meteorological data within various universities and departments, and a lot of the time they are downloading the same sort of data into seven different places, because seven different researchers are working on it, which is expensive, because you have got to pay to download that data and you have got to store it somewhere. We think there are opportunities in data management to have one copy stored somewhere or one copy made accessible so that everybody knows it is there and can share it. Does that make sense?

The CHAIR — Yes. I am thinking of examples in some research I have been doing recently. I might follow it up with you later.

Mr GROENEWEGEN — Okay. The reason that ANDS was established was that, as I said before, more data is being produced all the time. There are more and more large machines that can produce very large amounts of data, and there are larger and larger models being created around that data. There are now computers that can process large amounts of data to create those models, so there are a bunch of issues around that. While software and hardware are relatively speaking getting cheaper, what we call wetware — people — is getting more expensive in that people are time poor, they have certain amounts of stuff they want to do and they do not want to spend all their time managing data, so we want to try to simplify that process.

We are also hoping if we attack this problem relatively early in this cycle, we will solve problems that will be expensive to fix later on. If we can start people thinking about data management now rather than five years down the track, we may actually solve some problems that will be much more complicated to fix later on when something has been done badly.

We certainly think the more data that there is online or made publicly available the more that can be done, because we are not replicating the same research over and over again. We believe there will be many instances where data that is publicly available will answer questions that the original data was not trying to answer. Someone else can take that data, process it in a different way, combine it with something else and come up with an answer to something — a different problem. We think that is an aim worth encouraging. That is particularly the case as there is more cross-disciplinary research being done. Many of the boundaries that used to exist when people were just physicists or just chemists or whatever are starting to break down. Often we find that researchers are very familiar with the data and the field they started in, but they are not so familiar with other data that might be of use. I had an interesting meeting at Monash University recently with a bunch of mathematicians who are all now working in climatology, because to be created the climatology models need mathematicians. These are people who started life as mathematicians and who are now climatologists. They need to know about things that are not what they initially started out their research life doing.

Therefore we think certainly there is a value to sharing data and that open access to data makes that easier. An example that we often use is the Hubble space telescope, which is a US government-funded facility. One of the requirements of getting time on the Hubble telescope is that you make all of your data publicly available after six months; you cannot just hold what you have produced. What is happening is that only a limited number of people can get time on the telescope at any given time, but there are now more papers being written about the secondary data — that is, the stuff that has been made publicly available — than has actually been written by people who have access to the telescope. Because people are using the telescope and making that data publicly available, other researchers are able to do work that they would not otherwise have been able to do. If they had to wait their turn to be on the telescope, they would have been waiting for years, because they get five times more applications to use it than the time they can physically have on the telescope. This is a real growth area in the sense that there are efficiencies there for everybody and you do not need to build a second Hubble; you can just reuse the data from the first one.

We particularly see this as an example of what you could almost say is government data. It is not the classic government data, but we think the government produces, both at a federal and a state level, all sorts of data that is of value to researchers across the sector. By the same token, the research data that is being done in the universities or the other institutions and agencies should be made available back to the government to help it in policy making. The water data-type stuff is a classic example. The more data there is available about water, the more the government can do about water policy. Therefore we think there is a need to try to coordinate some of that. ANDS is very interested in working with government agencies, and we already have contacts with government agencies at Geoscience Australia and the ABS and the Bureau of Meteorology to try to see how we can connect up other university research data with the stuff that is being produced by those sorts of government agencies. Our method of doing this is that we have four service delivery programs. There is a document that has a little summary of that, a one-page type thing. Did you get that? Did I send that to you?

Mr KOOPS — No, I did not get that.

Mr GROENEWEGEN — It does not matter. We have four programs in which we are looking at developing frameworks. We are looking at talking to government and funding agencies about how we can encourage data management and data sharing — for instance, talking to bodies like the Australian Research Council, which funds a lot of university research, about trying to mandate that a certain amount of data becomes freely available if it is funded by the ARC, and at the present time those rules are not necessarily in place. We are looking at providing utilities, which we see as tools that are available for helping data management — for instance, creating persistent identifiers so that you can reliably cite a piece of data over the long term and other people can come and find it. We have a program which we call Seeding the Commons, which is about encouraging people to make their data publicly available, to make it discoverable, and also building capabilities, as I spoke about earlier — trying to have more people understand data management and also encouraging researchers to think about data management as a part of the research that they would do, in the same way that they would do discovery of resources and so on.

We are also funding some development activities through the National eResearch Architecture Task force, which is known as NeAT. One of the things we are very keen on doing is trying to enable discovery of research data. So we have a program in place which will mean that at least you can find out that that data exists, even if the data itself is not publicly available because the researcher may only be able to offer access in limited circumstances. There may be confidentiality; it may be that the data is dangerous to use unless it is put into context.

Of the issues that we are trying to address, one is obviously cultural change around the issues in research and that people traditionally have not thought very hard about data management. For most people, if they back up their C drive onto a DVD once every now and then, that is pretty much safe. We have heard all sorts of wonderful horror stories about people whose idea of a secure backup is putting all their DVDs in an air-conditioning vent in their office, because that way no-one can steal them. It is true. We also heard about a researcher who was told that off-site storage was good and so she backed up her hard drive onto a DVD every Friday and then mailed the DVD to her mother, who lived in another state, so that there was always a safe backup copy in another state.

Ms THOMSON — That is secure.

Mr GROENEWEGEN — It is secure, but it is probably not the most efficient way of doing things. I do not know what her mother was doing with all the DVDs — presumably putting them in a cardboard box somewhere in her garage.

Ms THOMSON — She could be doing some interesting research.

Mr GROENEWEGEN — She could well be — data sharing at its finest. We know that many researchers do not really see the value of data management, and that it is extra work that at this stage they are not necessarily wanting to do and that there are no obvious benefits — and the same goes for data sharing. You can understand philosophically why you might want to share something, but the ‘What’s in it for me?’ thing is always a difficult one to overcome. It is a very complex sector with a lot of different people involved in it. We do not have a huge number of staff, and we realise that trying to deal with any large number of researchers is always tricky. To make as much of this discoverable as possible it needs what we call metadata — descriptive data about the data. For that to work it needs to be easy to generate because no-one wants to write metadata; it is the most boring thing in the world to do. But if you do not do it well, no-one can find what you have done. We need to try to work around those sorts of issues. We need to engage the researchers into new types of technology, because we know that it is always difficult to learn a new tool unless it looks immediately obvious that it will make your life easier. And we need to get enough data out there to start building a sensible momentum, because we recognise that, again, until you can see that there are things that are of value to you, it is hard to know why you should do the same thing for someone else.

Finally, as I said before, we think there are some efficiencies to be gained in that a lot of things are rediscovered, a lot of experiments are re-run because people do not know that the same experiment has already been done. That can be very inefficient. One of the things we want to try to encourage in this space is the publication or the accessibility of failed experiments, because too often people will try to do the same thing that someone else has already found out does not work. If everybody could avoid re-doing stuff that someone else has already figured out does not work, we could concentrate on figuring out the stuff that does work. And in many areas there is a lot of failed work. We have been working very closely with crystallographers, who look at the make-up of proteins. To do that they have to be able to artificially create a protein, and that is a very tricky thing to do. Eighty per cent of the time they get it wrong: they try to build the protein and it fails or it does not crystallise properly or they cannot use the result of that. What generally happens is that they go through a process: they write down everything they want to do in their own little notebooks, or whatever, and say, ‘Well, we won’t do that again’ — and that is as far as that piece of information ever goes.

If another researcher is trying to do a similar experiment and trying to build the same protein, they have to go through the same process, because no-one ever shares failures. We think there needs to be a certain level of recognition that failures are in fact part of the research process and are part of the work to a certain extent, and that sometimes telling people about your failures actually makes people’s lives easier.

I'll go to some assumptions about what we are trying to do. As I said before, we know that not all data can or should be made available, and that this is not a blanket thing. You cannot say everything that is research data has to be made publicly available at such and such a day because there are a huge range of issues around confidentiality, copyrights, patents, national security and so on and so forth; that make that very complicated. It also means that it makes researchers nervous about this because for them the data is their way to publication, and publication is their way to promotion and success, so if you make them give away everything they start to wonder what they are doing it for. But we would at least like to encourage its existence being made known. Even if you do not necessarily give away the data, if there is some metadata that says, 'I have done this research, I have this dataset, I might be willing to share it under certain circumstances' that is still better than no-one knowing that it exists at all.

The other thing is, as I said before, ANDS is really only the beginning and a lot of what we are trying to do is just what we see as the beginning of a much larger process. We do not think we are going to solve this even in three years. There is going to be an ongoing process beyond that and we hope to have proven enough success in three years that we can then apply to the Federal Government for more funding. It is a partnership between CSIRO, Monash and the Australian National University and we are looking to grow that partnership further with other institutions and other research organisations; and we expect to learn as much as we expect to tell people in this process. We really do not think that we have all the answers yet, so we are going to be making mistakes and trying to figure out what will work better next.

The CHAIR — Looking at our terms of reference, if you have a copy of them handy, a lot of what we have to report on are the potential economic benefits and costs for maximising access to and use of government information for commercial and non-commercial purposes. Tying that in to our inquiry, would you like to link a little more of what you have outlined to what we need to address in our report.

Mr GROENEWEGEN — Yes, in terms of economic benefits and costs, from a basic level we would hope by making more data available we are creating efficiencies in research that otherwise is often funded out of the public purse. Most university research is paid for either by the state or federal governments through grants. We would expect that if you make more data available you will not have to pay to do the same research over and over again; you will do more effective research.

The CHAIR — Then my next question is: given the taxpayer has funded so much of that research, how do you get the government to examine the research and the data that you will be compiling. For example, when legislation is being framed there is an inordinate amount of data, information or research papers to back a particular policy option. There is also information that could back an alternative policy option and legislative option. I know this is not part of your brief, but it is part of our brief, and that is how we maximise the use of all this data, information and research that is directly or indirectly funded by the taxpayers to generate economic benefits for Victoria, for commercial as well as non-commercial purpose use?

You mentioned water. I was up in Canberra yesterday and there was debate about what the right approach is on water and it often does not go to the scientific analysis and the vast amount of university research on water; it gets down to more like political posturing. It is the same with contentious or non-contentious legislation. There is a significant amount of data and research in our universities that back particular ways of doing things, but the Victorian Government and Victorian MPs do not necessarily pick up on it, or follow the clear signposts.

Mr GROENEWEGEN — I do not know if there is anything we can do to make people follow clear signposts.

The CHAIR — Engaging the government, engaging politicians?

Mr GROENEWEGEN — Yes, it is an interesting question in a sense. I mean I understand what you are saying, but it is not what ANDS is trying to do. We are about trying to make researchers work better with each other, so I presume that the only way we can engage with politicians is to try and get them working together to a point where they can reach a consensus that they can put to a politician, and to hopefully bring together enough data that the clear signposts become clearer. My understanding is that the Bureau of Meteorology's National Water Information Service, which is coming soonish, is an attempt to do exactly that sort of thing — to pull together all the various pieces of water information from across the country. I think they have 260-odd sources of water data from across the country — various state governments, local councils — all sorts of places that collect data about water and pull that all together into a single place, so that if you need to know how much water there is in Victoria, there is only one place that you need to go.

Those are the sorts of things we would like to do more of. Some of this stuff I suspect is always going to be below the level at which anyone other than the researchers is going to make a decision about because it is about low level research that makes sense to other researchers and does not make sense to the rest of us. I read the stuff all the time and I do not understand it, but hopefully it will allow people to bring together more information to make some of those answers easier because at the moment you have to go all over the place to find all those bits of information. If there was more cooperation, more bringing together places into a single place and a single place to look for it, then at least hopefully you will get a chance to see more of the information rather than having to go everywhere to find it.

Mr CRISP — I have been absorbing your summary because I see that some of the issues and problems are the same ones we are grappling with as we try to obviously fit Victoria into an ANDS-type program. That standard for metadata is something we have been looking at, making a transition so that people put their stored information into a metadata form, and I take on board what you say about how thrilling we are going to find that, so that then it can be searched or made available. But will ANDS be setting standards for metadata repositories? Is that within your brief, because if everybody develops their own metadata standard, we have still got the same confusion that is holding back access to our information?

Mr GROENEWEGEN — What we are hoping to do and the way we are hoping to do it this is to — —

I think 'standard' is too strong a word because 'standard' implies something a bit more rigid than we think is actually going to apply. I think one of the troubles with trying to impose standards for metadata is that it means that you end up with either a huge list of stuff or stuff that no-one will ever fill in because they do not understand what you are asking for.

What we are hoping to be able to do is set a minimum of best practice metadata that we can use for harvesting purposes, so if you look at the diagram under the discovery services there you will see we have a whole bunch of different databases, data stores that we would like to be able to harvest from and that we expect to have to do some sort of normalising across. So that if your name is Christine Campbell in one repository and Chris Campbell in another one, we can figure out that that is the same person or it is not the same person, as the case may be. This is a persistent problem we have that researchers do not use the same name in the same place. In fact a lot of people do not use the same name. You have a name in the context of what you are doing, and that will vary, so we want to try and pull together a minimal amount of information. We will take pretty much any amount of data we can get, but we will assume that most of that is what we would call specialist metadata. We would still make that discoverable to a certain extent.

We really need to know stuff like who you are, what the name of your project is, what the name of the data is. As you can see from the little blobby bit, we want a collection, a research or a dataset. A name for those sorts of things is enough. We will draw the lines between those, and we will let the researchers sort out what metadata they need beyond that. No-one other than the researchers can sort out the specialised metadata in our experience — not even library cataloguers who have been dealing with this sort of stuff for decades. When we throw at them some of the stuff we get off the big machines — the synchrotrons and whatever — we do not know what to do with it. We do not know what is important and what is not important. We try to capture all of it, put it somewhere and say, 'Here it all is. Now that you know that dataset is there, you can look at the other metadata if you know what that means and you can decide whether that is of relevance to you or not'.

We are trying to put some of our tools in place to try to capture that sort of stuff. We want to capture as much of that automatically as we can. We want to have tools that say, 'If you log into this system, we know your name, we know your affiliation and we know your university, so we do not want you to type that in again'. We will pull that from some other part of the system and say, 'Here is your name, here is your affiliation and here is the thing that you are working on'. Where possible we would prompt them by using other bits of the system to say, 'Is this the project that this metadata relates to?'. Most universities at least keep track of where their grant funding comes from, so we will say, 'We know you have got an ARC grant for X, and we know who you are'. If we can then say, 'Professor X has got an ARC grant, and they want him to put some data into this repository. Is this the ARC grant that that data belongs to?', then we can draw that line and they do not have to type in, 'This is my such-and-such grant', which they probably cannot remember the name of anymore anyway, or they will type in a slightly different name. This was another trouble we had. If you let humans type in metadata, they do not do it consistently.

The CHAIR — Who types it in if it is not humans?

Mr GROENEWEGEN — This is why you want one source of a lot of this metadata and let the computer sort it and tie it together. If you make them type it in over and over again or you let them type in what they think is the right answer, it becomes inconsistent. We did a process a couple of years ago where the university was trying to find out what journal articles it had published over the last five years. We have a system that managed that. One of the systems that they had to put in was the International Standard Serial Number, which is an eight-digit number. The field that they could type that into did not say in what format you had to type that number: you could type it in as four digits, a hyphen and four digits; or four digits, a space and four digits; or as eight digits. The system did not stop you from doing any of that, but somewhere along the line we had to match all those numbers so that you knew that the same journal was the same journal. It took a lot of work to tidy that up — assuming that they had typed the eight digits correctly! It is a messy business. You do not want to get me started on that.

Ms THOMSON — Yes. It sounds like a big task. From this exercise you are going to end up with the virtual research lab where people will be able to get access to information, share information and potentially even work more collaboratively than currently occurs. Researchers might want to have access to initial data that is government data, and my issue is that how usable that data is. More importantly, as the standards are developed here by the Monash project and the work that CSIRO are doing collaboratively, the question is whether that will also be delivered to governments to look at the way in which they deliver data and information for those research projects and purposes.

Mr GROENEWEGEN — We think that there is lots of really good government data that we would like to incorporate into this sort of thing. We know from researchers that a lot of them rely on government data for all sorts of research. They would like to have that availability and connection. We would have to make what we were doing available to government, and it would need to work in with government systems as best it can. We see that as a key part of what we would like to do. We realise it is probably going to be tricky, but we know that if you ignore the government data, you are ignoring a huge section of what researchers need and want and stuff that is being done. Governments collect lots and lots of useful stuff that people could be using and, I suspect, do not even know of.

Ms THOMSON — The issue for us — because I think we would all like to see our researchers get access to that information — is whether it is better to allow the work that you are doing, the work that CSIRO and Monash are doing, to take place and to have our government data presented in a way which suits that process, or whether there is some way that those discussions — whether or not they are occurring now — between Monash, CSIRO and government about the way in which we provide information and the way we present data need to be looked at concurrently?

Mr GROENEWEGEN — We would like to discuss it with you, I suspect, because as I said before, we do not think we know all the answers and we do not want to impose standards on people that are not going to work. There is no point in us sitting there and saying, ‘This is the way everything has to work’, and then having it not working for anybody. We actually believe that we need to be in discussion with and to work with people. So we would be happy to work with the Victorian Government on how we can work out this system so that it works for both you and for researchers outside. There is no point in us just saying, ‘This is the way it will be’, because it will not be. We do not have any power over anybody to enforce that, and we know that researchers in general like to do things the way they like to do them. The phrase ‘herding cats’ comes up quite a lot when you talk to people about academics. So we would be very happy to engage in a dialogue about that sort of stuff.

The CHAIR — Thank you very much. We appreciate that, David. That pretty much covers what we need to know. You will be provided with a copy of the Hansard transcript in about a fortnight, and you are free to correct typographical errors, but you cannot change the substance of it. It will then be made available to people on the internet. Thank you very much.

Witness withdrew.