HOUSE OF COMMONS

Public Administration Committee

Oral evidence: Statistics and Open Data, HC 564

Tuesday 22 October 2013

Ordered by the House of Commons to be published on 22 October 2013.

Written evidence from witnesses:

> **Professor Sir Nigel Shadbolt,** University of Southampton and Chair, Open Data Institute and **Stephan Shakespeare,** Chief Executive, Yougov and Member of the Public Sector Transparency Board

Watch the meeting – http://www.parliamentlive.tv/Main/Player.aspx?meetingId=13989

Members present: Mr Bernard Jenkin, (Chair); Alun Cairns; Kelvin Hopkins

Questions 80-130

Q 80 Chair: I welcome our two witnesses to this session on open data and the 2012 Open Data White Paper. Could I ask each of you to identify yourselves for the record, please?

Sir Nigel Shadbolt: Professor Sir Nigel Shadbolt, the University of Southampton and the Open Data Institute.

Stephan Shakespeare: Stephan Shakespeare, YouGov and also on the Public Sector Transparency Board. I was chairman of the Data Strategy Board until it was dissolved into the Transparency Board.

Q81 Chair: The first question we would ask is has the Government understood what open data is really about? We get the impression of a great deal of enthusiasm, but have they understood the whole breadth and scope of potential for open data and are they acting on it?

Sir Nigel Shadbolt: They have a pretty good track record of pushing innovation in this space, but one has to say that there is no single reason for doing open data. People will talk about transparency and accountability; they will talk about efficiency in public-sector delivery; and they will talk about generating economic growth, all of which open data can be in the service of—or increasing participation or enhancing the quality of the data itself.

There are eight or nine distinct reasons, and that is one of the reasons that open data is also somewhat non-partisan. I saw a remarkable sight in the US at a meeting where the Tea Party sat down with the Democrats and agreed that data publication was a good thing.

The challenge for governments is not just to fix on a particular reason for doing it but to understand that there is this wealth and range of opportunities.

Stephan Shakespeare: There is good understanding across Government and there is obvious palpable desire to do the right things. You do not come across any significant defensiveness or obstruction to try to stop this agenda. It is widely supported, as Professor Shadbolt said, across the parties and across all the people I saw when I was doing my Shakespeare Review of data strategy.

The problem is that there is so much to do, it is so important and so urgent, and is Government capable of meeting the challenge in the way that it is looking at doing it? One of the things I recommended in my review was that there needed to be some external driver of it and that it was necessary to have a twin-track approach to data. I will not go into the details of this idea here, but it is a way of getting data out very fast and not leaving it in the hands of Government.

Q82 Chair: Mr Shakespeare, you say in your evidence that there is a need for a new strategy. Do we not have an open data strategy already? What is lacking from the strategy?

Stephan Shakespeare: We did have, and I am pleased to say I want to add to what I wrote then, because since then, on the Transparency Board I have seen two draft documents, one on capability strategy and one on creating a blueprint for an infrastructure, both of which are really good. There is no question that the Government response to the review has been positive and they have started to act upon it, and these two drafts, which obviously I cannot talk in detail about at the moment because they are not ready for that, both represent efforts to take this agenda very seriously and to put together a real plan. My worry is not that the plan is not good, but rather whether it can be driven at speed.

Q83 Chair: Professor Shadbolt, you said that the Government is setting its sights too low. What do you mean by that?

Sir Nigel Shadbolt: In some respects, the UK is world leading. It has put really rather difficult data sets out there at scale, everything from reported crime through to health data and spending data, but the "publish it and they will come" model is not quite enough. One of the reasons that we established the Open Data Institute was to generate a vibrant demand for open data. I do not think high-quality supply will flourish unless you have businesses and organisations and public-sector Departments whose business depends on the flow of that high-quality information. One area where the message is not fully understood yet is that the biggest beneficiary is the public sector itself, the Departments of State, in consuming their own information and the information of other Departments in a much less difficult, complex and expensive fashion. That is one of the real prizes of open data.

"Set too low" partly refers to the fact—and the previous evidence session covered some of this—that whenever you get into a discussion about whether we should sell the data or make it available as open data, too often the short-term interests of selling data for a few million pounds here and there trumps the wider economic and public good that would come from making it available as part of the national information infrastructure. Perhaps we will talk about that later. This is a very important development, this idea of a national information infrastructure. One of the most important things that we see coming out of

this development is that, in future, states like ours, countries like ours, will need to really take care that their national information infrastructure is supported, curated, maintained and enhanced in the same way that it is for roads, for electricity distribution or for power.

Q84 Chair: However, you say we risk being overtaken by other countries.

Sir Nigel Shadbolt: Yes, because this is such a fast-moving field. The field of data science—the field of what digital data does to make the business of administration, of Government, of participation more effective, more efficient and more productive—is moving so fast. Countries have looked to the open data examples of countries like the UK and the US, seen the benefits and opportunities and, in some cases, are not encumbered by existing structures such as trading funds or existing assumptions about what could or might or should be charged for or, indeed, are investing just more in the basic research and development and the human capability that you require to do this work.

Q85 Chair: Mr Shakespeare, you talk about the need for stronger, clear leadership, for driving implementation of a national data strategy, and say that there should be a single body with a single public interface for driving increased access to public-service information. Does that not already exist? Is that not what the body that is being created does?

Stephan Shakespeare: There is no single body that does that. We did have the Data Strategy Board that was sort of doing some of that and looking at it from the accountability side but much more the economic case—the business case. We have lots of people across Government all committed to the agenda and adding to it really well, but we do not have one author, one body, that is driving this in a joined-up way. It is notable, for example, that some of my agenda has gone into the Transparency Board, but the other bit has gone into BIS and its capability strategy. Both are extremely good responses, but I wish they were together.

Q86 Chair: We have the Science Minister and the Cabinet Office Minister, Nick Hurd, coming to give evidence together. What is the question I should be asking them about leadership? They are obviously going to be pretending they agree about everything, but if two Departments are running it, I can see there is always going to be a problem.

Stephan Shakespeare: Yes. You really do want cross-Department and cross-disciplinary co-operation, so that is the good bit.

Q87 Chair: Give me the evidence that there is not good cross-Departmental co-operation.

Stephan Shakespeare: How do you make sure these twin chariots go together? It is a difficult thing to achieve.

Q88 Chair: I ask the questions. What is the evidence that they are not working well together?

Stephan Shakespeare: I do not think there is evidence of that at the moment, because we have two strategies coming out that work well together and that are good side by side. I am simply predicting that it will be very difficult to do this quickly. The opportunity here

is huge, but it is a much bigger opportunity if done quickly. We need to have a speed-crazy approach to this—an incautious approach.

Q89 Chair: Why is there such an urgency?

Stephan Shakespeare: First of all, rather obviously, there are advantages, including cures for things and efficiency in providing health and services, that we are not getting if we are not using it now. Secondly, very importantly, on the economic side, I regard this as the second phase of the digital revolution, the first phase being just connectivity. Silicon Valley was the huge winner of that first phase. All the big companies we associate with the internet are centred on a very small area in California. There are good reasons for that, which include Government investment and leadership. People forget that the American Government had something to do with that; it was not merely entrepreneurs.

Q90 Chair: The Federal Government or the State Government?

Stephan Shakespeare: Both. There was huge support for creating a centre for basic science in this in Stanford and there are lots of ways the Government supports projects and things that become commercial projects later. In this case, we have here the most coherent, largest data sets, we have the expertise and we have a desire on the part of everybody to get this done. We could be the leader in the second phase. To be the leader, we have to be very urgent about it.

Sir Nigel Shadbolt: I would amplify that. Due to the size and the relatively homogeneous nature of the UK, we have a real opportunity to show just how data-driven delivery of both economic and social value can happen. The US is somewhat hamstrung by the fact that there is a large federal system; much of the valuable data lives inside states, and state law varies. Therefore, there is a real innovation opportunity.

I would also say when they come together, the challenge for those Ministers is to reconcile—and it was referred to again in the previous evidence—around the schizophrenia between whether we charge for the data, and many of the trading funds live in BIS at the moment, or make this part of the public task. Is this part of the public national infrastructure that is so important that it has much larger economic utility than being sold out of a particular Department's brief, run by people like the Shareholder Executive, for example? That is not to criticism them. They have been set up with a particular set of ordinances, and their incentives are aligned in such a way that they will look to monetise and sweat value from their data assets. However, there is a bigger picture here.

Q91 Chair: Can you just explain what the advantage is of making this data generally available?

Sir Nigel Shadbolt: If we take an example that was recently lost, in fact, when the Royal Mail was privatised, the postcode address file was sold as part of the privatisation. That previously was, potentially, a common good. It was sold even by the Royal Mail in public ownership, but the opportunity existed at that point for that data to be made available as open data—all the legal addresses in the UK. What is the advantage of that? Almost every conceivable new advance in delivery of services uses digital capability; everything happens somewhere, everything gets delivered somewhere, whether it is blue-

light services or commercial innovation. In fact, the experience was interesting. The Dutch had to buy their address file back.

Q92 Chair: How much do you think the postcode address file is worth?

Sir Nigel Shadbolt: Many hundreds of millions of pounds, potentially, to the wider economy. In fact, the Danes estimated the return on investment of making theirs publicly and openly available as up to 40 times what it is costing them to release it.

Q93 Chair: Do you think the Treasury is the ghost at the feast here? Obviously, we got a better price for the Post Office because the Treasury insisted that the postcodes should be included in the sale.

Sir Nigel Shadbolt: It is interesting to understand how material that data set was and whether we would have seen any substantial variation in that. Yes, the Treasury has, in the past, simply not been convinced or persuaded or had enough instruction in the fact that this is a new opportunity. I do not think they get much of the opportunity of the digitally disruptive economic abundances that can flow from data. That is still a relatively ill-understood area, but you name one or two examples and people go, "Oh yes, of course," and I will give you two examples. When the GPS signal was made publicly available, it generated a multi-billion dollar global market. The US has made its meteorological data available, and the secondary insurance market for weather data in the US itself is worth $8 billion. These are large amounts of value that flow from data available for everybody to innovate around.

Q94 Chair: Mr Shakespeare, you want to speed this up. Does that cost public money? With the support given to the birth of the internet and the digital age in California, it sounds like the Government was spending quite a lot of money.

Stephan Shakespeare: There are some things that you have to spend money on, which includes basic science, and I want to reiterate the importance of that, because it is something that commerce does not do. It does not invest in basic science. That is a legitimate role for Government—to support an industry by doing the basic science bit that companies are unlikely to do themselves.

However, the release of data in itself should not be something that is expensive, and I do not think it is. If you want to make it of the highest quality, there is an expense attached to that, and of course we recognise it, which is why I talked about the twin-track approach—that we should identify the data sets that need to be clean and need to be published to certain high standards, and that should be track one. All the rest is published as track two—quick and dirty, as one might say, so long as one knows that it is dirty—and left to the data scientists to do what they can. Obviously, there should then be a desire to move things from track two to track one as quickly as possible, but the fact that you cannot afford to do everything well and that you cannot do it at high speed should not mean that it is not available.

One of the reasons for making data available—and I would just like to say I strongly support what Sir Nigel said about geospatial data, for example, as being critically important—is that the expertise that comes to bear on it when you open up data is vastly increased. Ordnance Survey is a fabulous organisation—it is a really great institution that

we have, full of brilliant boffins—but they are not the sum total of intelligence that can be brought to bear on this. They are not the sum total of the creativity that can be brought to bear on this data and, therefore, the more open it is, the more that we can get public benefit out of it.

Q95 Kelvin Hopkins: You have touched quite a lot on some of the questions I was going to ask you about the address wars and geospatial data from Ordnance Survey, but isn't the privatisation of this information, where it has to be sold quite expensively, anti-democratic and certainly against the open-data principle?

Stephan Shakespeare: Obviously, I would prefer it to have been openly available and I have said that all along. However, I do not believe that it is absolutely essential that things are free in order to gain the value out of them. I do not want to make a big thing out of this, because it is a side issue, but the principle that you would normally apply is for things that are of value potentially to everybody, there is a very strong case for making them free. Obviously, geospatial data would fit right in there. Things that have a very specific value that may be costly to make available you could charge for, so I do not believe that all data necessarily must always be free, but in this case certainly it should be.

Sir Nigel Shadbolt: We can imagine the value-added services that people would sell out of the back of good data are certainly chargeable, and the Met Office is a good example. Its advanced climate-prediction models are sought after around the world and paid for. The underlying public task, the public weather service, is the piece that should be available and is available as open data now. It took some persuasion and time to get there, but we now do have that, and that is a good step along the way.

The bigger question is this is not a stationary object, and it is one of those areas where parliaments and governments in the future are going to be saying, "What part of our data estate needs to be held, in some sense, in public trust to generate the most value, economic, social, in terms of efficiency and so on?" It is likely that will not just be about releasing some data sets we already happen to collect. It will be about actively looking at the new kinds of information that we will need to run our countries efficiently and effectively. We are beginning to see this in areas like health service provision, and what we will need to connect well-being through to primary care through to secondary care. What will that world look like? It is going to look very different, because people will be generating so much more information from themselves about their state of health, their wellbeing and so on.

Therefore, the basic landscape as to what will comprise data that the state thinks has better value as a common good will be changing along the way too, and one is likely to see it in the new census that will be constructed, I suspect, going forward.

Q96 Kelvin Hopkins: The Government's attitude to a whole range of things that used to be in the state's hands has changed over time. My view has not changed, but Government's view has changed. We have a direction of travel now, with the postcode address file going, and other things may go; surely, in a meaningful democratic society, citizens must have a right to reliable information that is publicly accountable and guaranteed to be truthful, free and accessible to make judgments about society, who we vote for and so on.

Sir Nigel Shadbolt: I would agree, and it is probably in the interest of a Government to know what all the legal addresses in the country are and not have to pay a monopoly rent for that information.

Q97 Kelvin Hopkins: Is it possible that at some point in the future Government may realise they made a terrible mistake and bring it all back into public ownership?

Sir Nigel Shadbolt: As to that piece of data, I do not know. What people might start to think about is whether there are alternative ways of getting the same information in a different way. We are starting to see that starting to happen in many areas with the crowdsourcing of data that was historically the remit of a particular part of Government. Geospatial is a very interesting example, where the cost of doing very high resolution mapping is beginning to tumble and you are seeing start-ups—I know a number of them—where the proposition is that in a few years' time we will be able to supply huge amounts of this information from sources other than those they have historically been derived from. However, your question is a good one, and it is: what is the common infrastructure, the common platform of data, that is required for the greatest common, public good?

Q98 Chair: Can I interject? What you are basically saying is that it is not a left versus right argument. It is a question of whether information should be held by private sector monopolies or in monopolies, or whether it should be free. It might be a state monopoly, it might be a private monopoly, but it is whether it should be held in a monopoly and, therefore, what it generates is restricted, or it should be generally available so that everyone can build platforms from it.

Sir Nigel Shadbolt: That is a perfect characterisation. It is not left versus right. It is open versus closed, and what you get from open is open innovation, and that is the prize.

Q99 Chair: Basically, therefore, what privatisation of the postcode address file system represents is a closed, corporatist view of the economy as opposed to an open, innovative view of the economy.

Kelvin Hopkins: However, is the crucial difference that monopoly in private hands is driven by profit and exploitation, and monopoly in the public sector is publicly accountable—accountable to Parliament ultimately?

Chair: They are both destructive, though; that is the point.

Sir Nigel Shadbolt: What you have to ensure, if it is in public hands as well, is that it is driven out there to a standard and a level that is useful and, again, we have plenty of examples where the data is in public hands and is not out there yet.

Stephan Shakespeare: Another area where this problem arises is in outsourcing or partnerships with the private sector, because of course freedom of information and so forth does not apply in those cases. The Government has completely accepted a right to data; people have paid for the creation of this data, it belongs to them and they ought to have access to it of some kind. However, the moment a project is outsourced, the company has a right to claim commercial confidentiality, and that data, although it is public data paid for by the taxpayer, no longer fits into that scheme. As I suggested in my review, one very easy way that we could at least start to ameliorate this is, in procurement, to always have a box that says, "What is your open-data strategy?" so they are required to say in advance

what their attitude to this is. That could then make them feel that it may be detrimental to their getting the contract if they state that they will not share the data.

Sir Nigel Shadbolt: There is a stronger view that procurement should have a clause that says, "It shall be produced as open data."

Stephan Shakespeare: I do agree, yes.

Sir Nigel Shadbolt: There is a stronger version of that, and it is not about being anti-corporate. My whole mission for the economic and growth agenda of this, and what the Open Data Institute is dedicated to doing, is to allow the widest possible innovation to drive the largest amount of potential economic growth as well as the other goods that flow.

Q100 Kelvin Hopkins: In writing to us, Sir Nigel mentioned there is a capability gap and that the Government has a problem there; Mr Shakespeare mentioned that some of the data is not that good. It may just be that some of it is more difficult to assemble, more expensive to assemble, and the private sector, if it finds it difficult, will accumulate the easy data and sell it off, but the difficult stuff, which is expensive, it will not do. In the health service, they will not do difficult procedures; they will leave that to the public sector. They will do the cheap things, the tooth extractions and the ingrown toenails and all of that. The capability gap is seriously getting worse simply because the Government is squeezing the Civil Service and reducing staff, and we are going to have more problems in future. Is that not the case?

Sir Nigel Shadbolt: That is something to talk about when David Willetts is here. I sit on his Information Economy Council and am helping draft the data capability strategy and, again, they have made substantive investment in the area of big data, but much of this is in the area of hardware, which is important. However, the human skills you need to drive this are essential. Whether they are in the public or the private sector, there is a genuine recognition that we do not have enough of them and that we are going to have to do a substantial amount more to really take advantage of the inherent strength we have to exploit this emerging data market, so capability is key.

As I have mentioned in the past in evidence, within the Civil Service there is a real challenge. It is not just the well tried notion that it is largely PPE graduates, although that is part of the challenge. I can give you one or two role models within Departments who are producing extraordinarily high-quality data and others who wish to, and if I could simply clone some individuals about 100 times over, you would have a transformative effect. You do not need legions of these people, but every Department needs a small cohort of people who know how to do this. Four or five people across 15 Departments of State would make a huge difference in this space.

Q101 Kelvin Hopkins: We have seen the City over the last two or three decades sucking out a lot of the kind of talent that in the past might have gone into the public sector and it is now too expensive to buy back. My own view—and you seem to agree—is that outsourcing is causing a problem. However, you mentioned hardware, and we have had previous reports on Government IT problems, which have caused vast cost to the public sector—mistake after mistake after mistake, which has caused enormous sums to be lost to the public purse—and we have been given the run around by the private IT sector. As we have said before, do we not need a much bigger in-house IT capability to deal with all this?

Sir Nigel Shadbolt: You are developing quite an interesting one with efforts like the Government Digital Service. In working with systems integrators and the big IT companies, there are a number of facets of open that are important to make open work as an ethos, and it is not just open data; it is open licensing and open standards. In some respects, one of the worst sins that has occurred is the lock in to proprietary standards that companies are able to impose because they have a solution for this particular geospatial product or this particular accounting package, and it does not have any hope of interoperating with other systems.

Of course, their interest, quite naturally, is to build a somewhat closed environment here, whereas in the web we have been developing, the open standards that work there allow browsers to interoperate, and allow the content to flow. This is one of those kinds of geek questions where there are really well understood solutions to enabling basic exchange of information between systems that currently are siloed and locked away under their own proprietary formats. Again, it is right in the heart of procurement, so one of the big challenges is to understand how clauses around open data and open standards just become routine.

Q102 Alun Cairns: I want to pursue the issue of the presumption to publish. Do you accept that the noises from Government, to leave it as general as that for the moment, suggest that their actions are well meaning, are practical and they work?

Stephan Shakespeare: They are certainly well meaning, and I have nothing to suggest that they do not work. However, I repeat—and I really only have one tune on this, the question of speed—how can you get this thing done quickly? I do not think anybody is resisting the publication of data. Everybody accepts the general imperative to do that, but there are all sorts of reasons why you would be slow about doing it. One of them, of course, is legitimate concerns about privacy. Those concerns are usually overblown and, as often as not, used as an excuse not to do something, but there is a legitimate issue there. Some of this stuff is difficult, which is why I come back to the other strand of my tune: the fact that it is not perfect should not be a reason not to publish.

Q103 Alun Cairns: Do you accept that the prospective legislative moves will implement effectively a presumption to publish?

Stephan Shakespeare: I do not know.

Sir Nigel Shadbolt: What we have at the moment is a set of public data principles in the White Paper that are endorsed as Government policy. The question is whether they are being implemented routinely. A lot of people would think they are, but I suspect that when you dig deep you find that the first thing the permanent secretaries or the Department heads leap to say is not, "Have we published out those data sets that we have not published out yet?" The presumption to publish has some way to go. People think it is sufficiently difficult and challenging that you might need to legislate for it. When Tom Steinberg gave evidence, he talked about the whole issue around a right to data and extending, in a certain sense, presumptions around FOI. A good litmus test is to ask yourself, if you are going to publish data under an FOI request, why is it not being published already routinely? Bizarrely, for example, even if you do get the data out, you

do not have a perfect right to reuse it in any way you like, so there are limitations around that.

Q104 Alun Cairns: A Secretary of State once said to me that it was his ambition to be able to respond to every written Parliamentary Question by saying, "It is already available online." Is that a fair approach and an ambition that every Secretary of State should have?

Chair: It sounds evasive to me.

Sir Nigel Shadbolt: I have some evidence. We were at the last Transparency Board, and one of the Ministers of State said, "When I answer my Parliamentary Questions, my first port of call is data.gov.uk to see if there is any data on the fact," which is, at one level, very encouraging. However, is there, through the entire system, root and branch, from the bottom up, this notion that, by presuming to publish this stuff out, the business of Government will become more efficient, more productive and more effective? There are still a good few places where this is seen to be one of those burdensome, tiresome things you have to do on a quarterly basis. There is still more to do to communicate why, how and what the best practice is, and there are some great examples in Government.

Q105 Alun Cairns: Is that statement I gave, which came from a Secretary of State, seeking to avoid it by dumping so much data that it hides sensitive criticisms?

Sir Nigel Shadbolt: There have been accusations of hiding it in plain sight.

Q106 Alun Cairns: Or is it something that is seen to be positive, whereby it allows open, public scrutiny?

Sir Nigel Shadbolt: This brings us on to the interesting question about the interpretation of the data, which is where organisations like the ONS—Office for National Statistics—have such an important role: what is the narrative? What is the plausible interpretation? That is where the debate becomes so much more interesting, because people will contest how you interpret the data. That is the great thing about publishing it out there. The data is there, so it is how you account for the differences in mortality rates in hospitals across the week, or whatever the issue of the moment is. Putting the data out there is just part of the process, but it is an essential precondition, I would argue, and Government is trying, but to imagine the job is done because we have some public data principles is naïve.

Q107 Alun Cairns: Mr Shakespeare, do you accept that?

Stephan Shakespeare: Yes, absolutely. Both parts of the Government, in their statement when they began, committed to a right to data, and clearly everything has been based on a presumption that this is the case, and it may indeed need legislation to make it happen if there is no other way of driving it faster.

Q108 Chair: On this question of enshrining the right to data in the Freedom of Information Act, we have some witnesses telling us that this is essential. How much would you prefer that to be explicitly in legislation?

Stephan Shakespeare: I would like to see it enshrined explicitly in legislation. I think it is a fundamental thing for the future.

Sir Nigel Shadbolt: Likewise, because policies come and go; legislation has a way of sticking around.

Q109 Alun Cairns: Do you accept that in local government and every other Government agency as well as Whitehall?

Sir Nigel Shadbolt: Yes. Some of our most valuable data assets sit in the local government context, and, indeed, through all tiers of Government, we need to really start to think about what the equivalent commitment at the local government level could do for the country, and very substantial improvements. You can begin to see this happen in certain, if you like, leading councils where this assumption is being taken.

Q110 Alun Cairns: Can I ask Mr Shakespeare about accessibility, the twin-track approach you have referred to quite a lot, and the developers and the general public in terms of the users of the data? Can you give me an example of a key data set in each? You talked about the dirty data; I think that is the phrase you used. What would be acceptable in that area, and what would be completely unacceptable and should be refined in the issue of the Act?

Stephan Shakespeare: I am loth to do that, because I am not a data geek and I would not like to draw those specifics. The Chief Scientist wagged his finger at me very authoritatively, saying, "Never refer to it as quick and dirty"; "imperfect" was all that he would allow. Therefore, I do not want to do that, but I do think that we can trust in interested parties to do a lot of the work that we would otherwise not be able to do within Government so quickly. There are lots of people out there who would like to see this stuff published and use it, but in order to invest in it, they need to be assured that it will continue to flow. One of the really worrying things is that you might dump some data and then not update it. Therefore, bigger businesses especially are not prepared to invest in working on this kind of data until they know that it is going to keep flowing.

Q111 Alun Cairns: Do you not accept that that could provide a risk for manipulation, whereby the dirty data is published and then that is used as an excuse, and the uncomfortable data, if you like, is updated at a later stage and, therefore, just having that as an accepted principle allows for greater manipulation?

Stephan Shakespeare: That possibility is there, but if we are going to worry about all the things that could go wrong, we will not do any of it. This is part of my argument against worrying too much about privacy. We get very nervous about medical data being shared and that somebody might, using very clever techniques, find out that a particular person has a particular problem. That should not stop us from making it open. When I say "open", I do not mean openly accessible to anybody, but anybody who has reasonable credentials, whose use of it can be tracked, should have access to this, because the advantages of getting that data used to find cures, better health regimes and so on are enormous. The worry that somebody somewhere might have some medical record revealed is very distant, but it is not a huge problem anyway if it does happen.

Q112 Chair: Do you think there is a role for the UK Statistics Authority in stipulating how there should be a routine publication schedule of data sets, so that they come out like GDP

data or unemployment data and then industry and other users have predictability, which they do not have at the moment?

Sir Nigel Shadbolt: That is a very good point. It could be very helpful to have some kind of guarantee of supply and quality and the sorts of attributes that you have associated with the data. Data about the data, so-called metadata, is crucial here. Again, statisticians live and breathe this and understand a good amount about what guarantees the quality of a data set. Also, at the Open Data Institute, we have been working on what we call open data certificates. These are, essentially, encapsulations of best practice, and we would very much like to see, and Government is committing to, in some parts, its data sets being published against this level of, in some sense, independent scrutiny.

The other thing to say about this whole drive towards an information infrastructure is some of this stuff is not deeply difficult. The most basic things are: what is there, where is that stuff and how does it link, whether it is schools or companies or contracts? One of the things that convinces me we have not done the job yet is that some of our most basic, authoritative data around what there is—the list of companies Companies House keeps—or where it is—the postcodes—are not routinely used across Government to link their data together. We have standard ways now in which we can represent this information that would allow data linkage to be much more powerful, but it is not routine in the way the information is published out. Therefore, there is a whole piece of work to do to increase our capability to publish the very good advances in open data as good quality open data.

Q113 Chair: Which are the better Departments and which are the less good Departments?

Sir Nigel Shadbolt: It would be invidious to draw up a league table. You can get a sense of that by just looking at who is publishing what, although some Departments are endowed with more data than others, perhaps.

Chair: Help us to have a sense.

Sir Nigel Shadbolt: I can give you a couple of examples of really good practice. The Department for Communities and Local Government (DCLG) has a site that is comparing and interlinking data statistics from across all local authorities. That is a very good example, sufficiently well done that the Department is starting to consume its own representation of this data in lots of other reporting that it is doing. The recent Public Health England site Longer Lives is an assembly of information about mortality and disease rates that is very accessible to the public and is a really good example where open data has been brought to life and begun a very interesting debate around variations in disease rates, death rates, mortality rates, up and down the country.

Q114 Chair: Is the Treasury a model Department on open data?

Sir Nigel Shadbolt: It has yet to be converted. It needs that Damascene experience.

Q115 Chair: Thank you; very helpful. Moving on to encouraging innovation, Mr Shakespeare, your review commissioned Deloitte to try to quantify the economic benefit of open data, and it came up with the staggering figure of £6.8 billion a year, comprising £1.8 billion of direct economic benefits and £5 billion of indirect benefits. Can you describe

why you think those figures are important and what the difference is between direct and indirect?

Stephan Shakespeare: Why they are important or why they are true?

Chair: Are they true? How true are they?

Stephan Shakespeare: First of all, the instructions that we gave to Deloitte were to only include things that the Treasury would be likely to accept as true, so it is a conservative figure, in our view. We said this is there to convince them, ultimately, and it needs to pass muster with them, so these are things that you can point to very concretely.

Chair: Are these annual figures?

Stephan Shakespeare: No. This is over a period of time.

Chair: Okay, fine.

Stephan Shakespeare: Taking an example that we all know, there has been a huge increase in open data and useable data about our transport system. It is something that we now take for granted, and we do not think of this as part of an open-data revolution, but you can find out where the train is that you are waiting to pick someone up from, or whether the bus is about to come around the corner. You can find out a huge amount of information about that. That data is obviously useful to the running of the service and so forth, but the indirect benefit is the amount of time saved by not having to wait for late buses and trains and so forth. There is a huge general economic gain from that data being made useable.

Chair: The larger figure is obviously of a more indirect benefit, but can you explain that?

Stephan Shakespeare: We give the example of transport in that; that is quantifying that. That is a number that then just balloons as soon as you start putting any imagination at all into where open data will lead us.

Q116 Chair: What does the Government need to do to realise this economic benefit?

Stephan Shakespeare: It is the same thing as I have said before. It needs to open it up. It needs to take responsibility for some of the data, and the bits that it cannot take such responsibility for it needs to open up more widely. It is the owner of this data. It has a huge responsibility for this.

Q117 Chair: Why do you not think the Government is doing this?

Stephan Shakespeare: They are trying to do it.

Chair: What is stopping them? Why are they only trying? What are the obstacles?

Stephan Shakespeare: The obstacle is the fact that it relies upon bureaucratic systems to do it. We talked about a right to data; legislation, of course, would force everybody in the bureaucracy to do that very quickly.

Chair: Who are the people in the bureaucracy stopping it?

Stephan Shakespeare: As I have indicated before, I do not think people are stopping it. There are all sorts of institutional reasons for doing it slowly: concerns about privacy, concerns about whether this is ready for publication, how awful if there is a mistake in it and we are discovered to have released a mistake. These are things that people worry about.

Chair: It is a cultural inertia, therefore.

Stephan Shakespeare: Yes.

Chair: That rings very consonant with our paper on reform of the Civil Service.

Q118 Kelvin Hopkins: Is there a capacity problem as well? As I mentioned earlier, if you strip out half the Civil Service and half the public sector, you are going to finish up with problems.

Stephan Shakespeare: There is a capacity problem no matter what you do, because there is so much of it, which is why we have to rely on internal expertise and external expertise.

Q119 Chair: In a nutshell, what are the three top things the Minister should decide to do tomorrow to make it all happen more quickly?

Stephan Shakespeare: I just have one thing, which is the things that they are doing driven on a day-to-day basis, not on a meeting-to-meeting basis.

Chair: By a single authority.

Stephan Shakespeare: By a single authority.

Chair: Okay. We have already discussed privacy a little bit.

Q120 Alun Cairns: Yes, I want to pursue it a little bit further, Mr Shakespeare. You have talked about sharing medical data and how we ensure that it is not tracked back to individuals' conditions and so on. How do you think we can reduce the risks in those situations? There is considerable anxiety out there about the sharing of data. On that basis, what practical steps can the Government take to reassure those who are most anxious?

Stephan Shakespeare: First of all, I want to make it quite clear that the revealing of personal medical data could be extremely painful to the person and that it is incredibly important to avoid that. I am arguing that it is extremely unlikely to happen if we do this right. We have what is called safe-haven technology, which means you can make data available in a way that you cannot take it out of the box, if you like, and you can access it remotely without removing it from the database. You can use it inside the box. Professor Shadbolt will be much better at explaining this, being a genuine data scientist, as opposed to me, who is absolutely not, but the data can be used inside the system; you cannot take it outside. Very importantly, you know who is using what. Therefore, if somebody is using data in a particular way, which then appears outside and is revealed later on to have been misused, you could put the responsibility back to that person.

What is incredibly important is that we think about the misuse of data and punishing that, rather than preventing any possibility of misuse of data. When we talk about data, we get

so worked up about it that we forget that the traditional means of safeguarding data are incredibly porous. It takes probably a few quid from a newspaper—maybe less so now—to somebody standing near a filing cabinet to get important and private data from where it is supposedly securely stored into the hands of a journalist. Why do we think that traditional data can be so porous and yet we must have no possibility of misuse from digitised data?

Sir Nigel Shadbolt: The challenge of privacy is crucial for us to address. There is a lot of confusion around what people are claiming for open data in this space. Open data, when it was about clearly non-personal information—where the bus stops were, what the contracts where—was not so controversial. When you start to look at the information-sharing schemes across aggregate data or individual data, there you have to say this is personal information and, where it is about the person, we have to use all means at our disposal, including existing legislation. The Information Commissioner's Office has powers so that, if people de-anonymise data, they can be held accountable for it. There is a range of things that we can do.

Fundamentally, when we look at the area of personal information, we have to do something rather bold, which is to rectify the deep asymmetry that currently exists between citizen and state, consumer and business. In fact, we all experience it, business knows loads about us; we do not have that same insight about ourselves that they have. The state has huge amounts of information about us; we do not have access to that information. There is a notion of really seeing—and this will happen because the technology is going to ensure it—a generation who become their own data controllers, to some extent. The expectation and assumption is that you will be in possession of your medical records to a certain level, your educational history and transcript, and your entitlements for work and pensions. If we can engage in the conversation in terms of empowerment back, it is a very different conversation from, "The state will work out how it is going to decide how to share information on you." Where that conversation has happened, in Scotland, in the NHS, there has been a much more interesting conversation around informed consent, with the result of very little opt out of patients who are seeking to share their social and health data together.

Q121 Alun Cairns: Do you think that technology can play a bigger part in terms of protecting individuals? Technology can be developed to effectively anonymise the individuals concerned, but it does allow for the aggregation of various data to be brought together to bring real benefits for the sharing of information.

Sir Nigel Shadbolt: There is, in fact, a lot of work under way on anonymisation, but it is not a guarantee and it is very important to understand the limits of technology, and it is very important to understand how we can provide accountability and redress should information be decrypted in various ways.

Q122 Alun Cairns: Do you think the anxiety issues are simply insurmountable in some quarters, or do you think there is some common ground?

Stephan Shakespeare: I have tried to construct a case where somebody would deliberately cause a big problem with data in this way; it is hard to imagine that this would happen. Occasionally, when privacy campaigners have tried to show that it can be corrupted, they

have managed to reveal some tiny detail about somebody very prominent, because that is one of the ways that you can work anonymised data out, because it was known publicly where somebody was.

I just cannot conceive of a likely scenario in which there is a serious problem here, and we do have legislation that can punish offenders. That should be tightened up and it should be used much more. That would be a more effective way of safeguarding privacy.

Sir Nigel Shadbolt: It is an existential issue that we have to face. Do not expect the technology to be able to solve the problem for us here. Often what will happen is that it is a set of data that is not released by Government sitting alongside other data that allows you to decrypt, and often this is volunteered data by the individual themselves.

A famous example is a set of data that was released by an online entertainment company about the films that people watch. They were trying to get people to provide more innovative solutions around recommendation systems—systems that would recommend films to you. This anonymised data was released, and then people discovered that there was a site where people had been rating exactly all the films they had watched; it is quite a well known site. You could take that data plus the other data and then find out exactly who the person was in the released anonymised data and, of course, they had been watching stuff that perhaps they were not so keen to have shared around the place.

Therefore, there are always issues here, and it is about an informed discussion around the benefits. There is no one so concerned to share their data as somebody suffering the difficulty of a particular disease or health condition. At the point at which it becomes very much in their interest, they are very much more motivated and incentivised to share the information to benefit themselves.

Q123 Chair: How much do you think this reticence and anxiety about privacy is based on a false understanding of what people are really worried about?

Sir Nigel Shadbolt: Could you elaborate on that?

Chair: How much is the new digital generation not very concerned? They put all their personal data on Facebook.

Sir Nigel Shadbolt: This is always said.

Chair: What would have been, in an Edwardian age, a matter of deep anxiety for it to be private is now something we just splurge all over the internet anyway.

Sir Nigel Shadbolt: Privacy is a human right, and we should all be hugely concerned to protect and defend it, not least because it allows for autonomy, not least in the political space. I can have some guarantee that my thoughts might just be my own and those of the people I confide in. The so-called generation that is giving up on privacy, interestingly enough, have a very nuanced view of what is available to open and what is not. As they grow up—I have seen this process—from no concern at all to a recognition that this will stay with them in their interview process as they go for jobs in the future, they become much more concerned about the issues and limits of privacy. It is not dead as a concept, despite what many people would claim, and it is in all of our interests to keep the reasonable expectation to privacy high and in place.

Chair: Mr Shakespeare, do you agree with that?

Stephan Shakespeare: Of course I agree that it is important to keep expectations of privacy high, but I do not think you can have both an absolute right to privacy, which I am sure you do not think there is, and an absolute right to data. The two things are bound to conflict.

Sir Nigel Shadbolt: There will be a tension.

Stephan Shakespeare: I think what you are referring to is when there is a newspaper report about the leak of someone's information: it causes great consternation and the anecdotal stories become big issues suddenly—someone has lost a disk and so forth. In reality, though, the moment they have sounded off about how terrible that is, they share vast amounts of private information that is highly leakable to Facebook and all the rest of it. In fact, the whole cookies regime that the EU created, whereby every site has to say, "We are going to leave cookies here," has been an enormous waste of time. It has not made any difference at all to people's usage of the internet. People are quite prepared to be tracked if it gives them a slight increase in the speed with which they can get information, it turns out.

Sir Nigel Shadbolt: There has also been a recognition that, in general, the presumption has been to pull out the privacy argument as a good reason not to do stuff, and that we should not allow people to hide behind the skirts of that without a good discussion about the benefits and disbenefits.

Q124 Chair: Very briefly, how much do you think an opt-in model would be a good idea? For example, could we all tick a box on our medical records saying, "I want my medical records to be used for the benefit of science"?

Stephan Shakespeare: I would definitely prefer an opt-out model. I think people should have the right to opt out, but they should make the effort to do that.

Q125 Kelvin Hopkins: I just want to make a point, following your line of questioning, Chair. There is a difference between having one's details anonymously used for a statistical survey in the health service and you personally letting everybody know you have a bad ingrowing toenail.

Stephan Shakespeare: Yes, but the point here is that the data is always going to be made useable in an anonymous form. However, because it needs to be case-level in order to be valuable, there has to be a risk that somebody who was determined to reveal who they were could do so. I say you deal with that. This is a rather far-fetched analogy, but we do not say to house builders, "You cannot build a house unless you make it burglar-proof." We say, "If you burgle, the police will come and take you and lock you up."

Q126 Kelvin Hopkins: There is also a difference between choosing to put all your personal information on Facebook, which I have chosen not to do—I do not like all that sort of stuff—and having it forcibly put on to the internet against your will. There is a distinction.

Stephan Shakespeare: Nobody is suggesting that you put personal information forcibly on the internet.

Q127 Kelvin Hopkins: However, the Chair is saying we are too fussy about privacy and we ought not to worry about it.

Sir Nigel Shadbolt: We need to make the benefits case more strongly. That communication needs to happen. I do not think we want to find ourselves in moral panics around unintended releases. Stephan is right; we do absolutely have to realise, ultimately, since we cannot guarantee that there will not be other data disclosures that will allow you to use that as the Rosetta stone to decrypt this information, that we should have strong forms of redress and accountability.

Q128 Chair: Thank you very much indeed. Can I just ask for a parting thought? How optimistic is each of you that the Government is going to get on the right track and the United Kingdom is going to be the world leader in open data and the open-data economy and the economic growth and opportunities that arise from that, on a scale of one to 10?

Stephan Shakespeare: There is almost no disagreement about this, and therefore I have good reasons to be optimistic. We already are a world leader and will continue to be, so probably a seven.

Sir Nigel Shadbolt: We have so many favourable initial conditions here, but there are clear dangers, and they are that we fail to convince and show and illustrate, and we have lots and lots of examples, where the wider economic value is served by releasing this stuff; and that we rely on principles rather than some firm legislation, which could really make a difference here. We need to understand that open Government data 1.0 is maybe happening, but the really interesting question is what does open data 2.0 look like? What happens when we really do understand how to get people participating back in the process of providing data to Government and how would that change our public services? All the prizes are there to be won. The UK is well positioned. I really think it would be fatal to imagine the job is done.

Q129 Chair: How much is your anxiety about a failure of imagination—

Sir Nigel Shadbolt: That would be a very good way to put it.

Chair: —rather than a failure of leadership?

Sir Nigel Shadbolt: Yes. We would do well to reaffirm those principles on a pretty frequent basis: that we are still convinced and trying to do harder in data release. That needs to come from the top political leadership, as it has under two Governments now. The remarkable thing is that this can be non-partisan and it has had very high-level support. We should keep pressing on that and reminding people of the benefits, because as Stephan said, there is virtually no disagreement on the potential benefits, and yet we find there are these frictions in the way, and perhaps that is inevitable. One of the good things about a set of national competitions in this area is that a race to the top can help, so we can look to other best practice.

Q130 Chair: Mr Shakespeare, the last word.

Stephan Shakespeare: You asked me not to ask questions, but I am going to ask a rhetorical question. Who is the leader—who is the accountable person in Government for data strategy?

Chair: We are going to see two of them in front of the Committee at our next session. Thank you very much indeed.