ANNEX

CACI Limited Responses to the Consultation's questions

Questions for consultation (page 6)

1. *Do the definitions of the key terms (on page 5) go far enough or too far?*

The definitions work well.

We consider it important that a Dataset should not be required to have any associated Information (interpretation and analysis) under these definitions.

We agree that private sector providers who have been commissioned or funded to provide a service in relation to public data should be included in the definition of public services. Examples are given below of barriers to free data dissemination resulting from private sector organisations being commissioned to provide data to the public sector.

2. *Where a decision is being taken about whether to make a dataset open, what tests should be applied?*

There should be a presumption that all raw data should be made public in basic text formats wherever this is consistent with the protection of personal data.

Where it is deemed necessary to carry out processing, formatting or provide interpretation of analysis of the raw data the cost should be balanced with potential use and benefits to the economy.

3. *If the costs to publish or release data are not judged to represent value for money, to what extent should the requestor be required to pay for public services data, and under what circumstances?*

In principle, we think that data within the public task should be Open Data and should be free.

If an existing dataset is considered to be of interest to a small user base it seems appropriate to charge for the marginal cost of processing involved with its supply.

If a data set exists, but users wish an extract, analysis, or other processing of the data that does not currently exist, it seems appropriate to charge for the marginal cost of processing involved with its supply. (An example of this might be re-aggregation of an administrative data set to different geographic areas.)

If a dataset does not exist in any form we do not think the principle of Open Data should require that it be collected.

4. How do we get the right balance in relation to the range of organisations (providers of public services) our policy proposals apply to? What threshold would be appropriate

to determine the range of public services in scope and what key criteria should inform this?

We suggest the range of organisations should be quite wide. There seems little logic in too much selectivity, for example saying that health data should be open but information on public leisure facilities should not be open.

Similarly the range might be wide in terms of the size of organisation. For example Local Authorities have datasets of house building completions and planning applications that offer potential for creating value added products and services that might stimulate economic growth.

It may be that they adopt different formats of Dataset and that some hold more Information than others relating to the dataset. However if the data is open there are opportunities for innovative products and services locally. Moreover the private sector may take the opportunity to combine the disparate data sets to create a national product or service and so more widely stimulate the economy.

Open data should not be restricted to national datasets or that which is common across the entire UK. Private sector users are used to different data, data formats, and even licence arrangements for separate parts of the UK. We would prefer more open data rather than a restricted set of neat clean data.

5. What would be appropriate mechanisms to encourage or ensure publication of data by public service providers?

It is important to bear in mind that a reasonable amount of data is already published by public service providers. CACI has experience over 30 years of developing private sector products and services based on such data. Over that time we have seen increasing openness and constructive engagement by public sector organisations. We believe this trend needs to continue at a greater pace. In many instances this might be encouraged rather than coerced.

Our experience is that when the public sector feels that it is their job to make data available they will actively do so.

We would welcome a clear statement indicating that

- Publication is the default position (for all data except for personal data.)
- All organisations have equal rights to access data.
- Data made available to the general public is also available to businesses.
- Publication should not be delayed or avoided by reason of the (lack of) quality of the data

In addition such a statement should

- Clarify the priorities should resources limit what can be done immediately.

- Clarify that *when suitably processed*, for example into aggregate statistics, data may no longer be personal (*subject to appropriate checks) and yet may be of much greater value*.
- Clarify that the Information Commissioner's responsibilities will include oversight of the above checks.

Some specific examples illustrate why we feel these are important.

1. *Data made available to the general public is also available to businesses.*

   The Valuation Office Agency has for many years allowed members of the public to look up the council tax band of any address in England. A decade ago this entire database was, in breach of copyright, 'scraped' from the web site and offered for sale to a number of private sector companies. Preferring to obtain data legitimately CACI informed the VOA and asked if it might be possible to purchase a copy of the database published online. This request was refused.

2. *All organisations have equal rights of access to data.*

   Academics often pitch for business in competition with CACI, and sometimes have preferential access to data. Academics are sometimes perceived in the public sector to be 'one of us' and not subject to barriers experienced by the private sector.

   In one instance the DfES allowed an academic to analyse the PLASC school pupil database – using value added products from both CACI and one of its competitors. The academic, who had worked for the competitor in the past, published data allowing the competitor to develop value added products based on the PLASC data set. When CACI requested equal treatment, or alternatively access under the same conditions as the academic, this was refused.

3. *When suitably processed, for example into aggregate statistics, data may no longer be personal.*

   CACI is currently in the process of obtaining access to the CORE dataset from the Tenant Services Authority. This database hold records for individual addresses and in the past it is quite likely that any request would have been refused on the grounds of data protection.

   Constructive dialogue has led the TSA actively engaging with the possibility of supplying anonymised data restricted to a limited set of data variables to the private sector, with the fees for the processing being borne by the requestor.

4. *When suitably processed data may no longer be personal, and yet of greater value.*

   For many years CACI has purchased from ONS a version of the Expenditure and Food Survey (now called the Living Costs and Food survey) enhanced by the addition of private sector information to the survey responses. With ONS' increasing emphasis on avoiding the disclosure of personal information in the early part of the last decade meant this process required approval by a new disclosure committee.

Since CACI were not aware of this committee ONS staff volunteered to assist by preparing the application for continuation of the access that had occurred for many years. However not being fully conversant with the CACI information appended to the survey they were unable to automatically gain approval.

Their response was to open a detailed dialogue with CACI, introduce CACI to the specialist disclosure statisticians and to agree a process that maintained data privacy to the satisfaction of the disclosure committee.

Policy Challenge Questions (Section 8, page 22 onwards)

An Enhanced Right to Data Questions (Page 25)

1. How would we establish a stronger presumption in favour of publication than that which currently exists?

- Establish the principle that data is by default open in existing and new legislation.
- Ministers and senior managers stating that their presumption is that data will be open and so engendering an open culture within their departments.
- Possible enforcement via the Information Commissioner.

2. Is providing an independent body, such as the Information Commissioner, with enhanced powers and scope the most effective option for safeguarding a right to access and a right to data?

The Information Commissioner with enhanced power and clear independence will give the public confidence that 'open data' does not mean less privacy. This may also be a good option in the context of open data promoting transparency and ensuring citizens have access to their own data.

However the IC's role is regulating proper use of data and this may easily conflict with promoting increased publication of data.

To safeguard a right to access one option might be the UKSA. Since their role covers an interest in the use of data and statistics they are a natural choice, at least in terms of access for organisations wishing to create value-added services.

3. Are existing safeguards to protect personal data and privacy measures adequate to regulate the Open Data agenda?

Yes existing safeguards are sufficient and are applied very thoroughly. In some cases government offices might be thought to have been too cautious, focusing

on small or obscure theoretical risks of disclosure whilst giving much less weight to utility, benefits and value. There is a case for instead considering the practical (rather than theoretical) risk of disclosure.

4. What might the resource implications of an enhanced right to data be for those bodies within its scope? How do we ensure that any additional burden is proportionate to this aim?

We assume that the right to data is limited to data that is already collected, perhaps including data in a format not currently suitable for publication. It would be impractical if the right extends to data not currently being collected!

The right to data does extend to a right to have the data processed by the public sector into any format that one wishes. While processing, formatting, creating metadata and any other documentation all require resource raw data merely requires copying. Metadata and documentation may not exist in an elegant format, but our experience of public sector data is that the metadata is usually more than sufficient.

Businesses seeking Open Data likely to assist in stimulating the economy seem less likely to create a disproportionate burden. It may be another matter where Open Data is sought be the average member of the public. A requirement to process the data into an easy to use format and/or provide supporting information to put the data in context and give the layman the understanding of an expert might well place a disproportionate additional burden. It will depend on the data set.

Specific examples where there are little or no resource implications include;

- Data on council tax bands can be accessed for any individual address on the VOA web site. Data aggregated to output area level is published as a data file. However the web site must have a data file listing every individual address and its council tax band. Therefore providing more detailed information file is a simple matter of copying this file.
- The same is true for the VOA files of valuation list of business properties. The data may be messy but there is no burden in publishing it as it is.
- There are many instances of public sector data being aggregated to Lower-level Super Output Areas (LSOA) and published, for example on the ONS web site. There seems little additional burden in also publishing these data files to the more detailed Census Output Area level. If this more detailed level of publication does not violate disclosure controls (such as those applied during publication of census data at this detailed level) it will be of greater value than the coarser LSOA data.
- The 2011 Census Quality Assurance and the Beyond 2011 project involves extending such aggregation to a larger list of administrative sources. Again it is difficult to envisage any significant burden in immediate publication of these data files at census output area level. (Again assuming publication does not violate disclosure controls such as those applied during publication of census data at this detailed level.)

5. How will we ensure that Open Data standards are embedded in new ICT contracts?

Make the contract lawyers and negotiators aware of clearly stated aims, requirements and standards and state that these must be included in ICT contracts.

Setting Open Data standards (Page 28)

1. What is the best way to achieve compliance on high and common standards to allow usability and interoperability?

A Code of Practice based on the Public Data Principles provides an excellent basis.

We believe the Public Data Principles should explicitly cover the issue of merging, combining and processing the public sector data with other data *in ways that ensure no personal data is shared with third parties.* We have direct experience of constructive assistance from the public sector in achieving this, and the resulting data sets are of significantly greater value as a result.

We would place less emphasis on "high standards" in relation to the quality of the data. When seeking to create innovative new product and services that may stimulate economic growth much benefit can be gained from raw data, partial data sets, 'dirty' and imperfect data sets if they can be provided in a readable form with the minimum of metadata sufficient to allow interpretation.

2. Is there a role for government to establish consistent standards for collecting user experience across public services?

There may be some need to establish standards in the case of private individuals seeking their own data from public sector files.

More generally standards for collecting user experience seems both unnecessary and tricky. The sheer breadth of users makes it unlikely that user experience can be interpreted consistently.

For example, as an established private sector data re-user we would expect an entirely different user experience to that of the concerned citizen or business developer new to public sector data. It is not clear that the respective experiences would be best measured in a common manner or that consistent standards would be meaningful to us.

Consistent standards may be nice to have but may delay publication and beyond some very minimal metadata are not a priority use of resources.

3. Should we consider a scheme for accreditation of information intermediaries, and if so how might that best work?

There already are some such accreditation schemes.

For example intermediaries assessed by ONS as Approved Suppliers of census data and value added information derived from the census. However there are also many organisations that are Licenced Distributors of the Census. It is not clear that the accreditation provides benefit to either the intermediaries or the end users. It is also unclear that the performance of the intermediaries is of any higher standard than if there were no accreditation scheme.

For a scheme for accreditation of information intermediaries to be effective it would seem that;

- There must be performance standards applied to the intermediaries
- Intermediaries that do not make the grade must have their accreditation withdrawn
- There must be investment in creating and maintaining a scheme to measure the performance of the intermediaries
- The accreditation scheme must be sufficiently well known amongst the user base for the withdrawal of accreditation to be meaningful

After all Open Data principles imply that intermediaries will be able to access data whether they are accredited or not. The alternative is that only intermediaries approved by the public sector may access public sector data.

We do not consider an accreditation scheme to be a priority. Indeed accreditation works against the principles of open information. If accreditation were required for intermediaries the government or public sector is in effect controlling who is permitted access to 'open' data.

Corporate and personal responsibility (Page 30)

1. How would we ensure that public service providers in their day to day decision-making honour a commitment to Open Data, while respecting privacy and security considerations.

This is a question of the culture within the public sector organisations. Will strengthening and broadening the Public Sector Transparency Board mean that individuals develop an instinctively "open" response to requests for data?

As an illustration: CACI recently desired access to data on care homes. Suitable data was found on a web site but the presentation format meant that significant (manual) processing would be necessary to turn the data into a useable database. We made a simple email request asking if it might be possible to access the database used to create the web pages since this would save us some effort. In response we received an acknowledgement of our FoI request. Various emails

relating to the process of FoI requests were received over the following three weeks and eventually we received an email detailing all the legal ramifications relating to the release of the data re requested.

Appended to this large document was a helpful note and a link to a different web site where the data we wished was available.

That helpful note could have been sent in response in the first instance but the request (which was nothing to do with FoI) was routed into a complex FoI process causing pointless delay.

A culture avoiding the reflex routing of simple requests into complex processes, and allowing public sector employees to follow their instinct to be helpful would have produced a significantly better result. Too many more policies, management processes, and corporate responsibilities and the like may simply get in the way.

It may be that governance and regulatory models, transparency boards, introducing corporate responsibilities will engender such a culture. Equally praising the work of those parts of the public sector already providing good access to data might win hearts and minds.

2. What could personal responsibility at Board-level do to ensure the right to data is being met include? Should the same person be responsible for ensuring that personal data is properly protected and that privacy issues are met?

While personal responsibility may be a good thing the 'right to data' principle will be met by good management creating the correct culture within the organisation.

It seems logical to have one senior person responsible for both confidentiality and data sharing.

3. Would we need to have a sanctions framework to enforce a right to data?

A sanctions framework may be useful, particularly where the objective is increased transparency.

If the objective is stimulating economic growth we believe promoting a culture in which the public sector that is comfortable with providing data will prove more effective than applying some form of enforcement regime.

When CACI seeks access to public sector data it is not unusual for our initial contact to be treated with some caution. An important part of our gaining access to data over the past decades has been our establishing a relationship where the relevant people realise we appreciate their concerns (for example over data protection) and actively seek to ensure our use complies with their concerns. For those who ask in an appropriate spirit there is already a great deal of Open Data in the public sector.

We have always believed that aggressively demanding access rights to free data and/or demanding public sector resources be diverted from the public task may receive a less constructive response. Those who do so may, incorrectly, feel there is currently little Open Data.

In our experience innovative uses of data can involve processing of data sets, perhaps including merging with additional data sets. Goodwill is necessary for the cooperation required to find ways of achieve this while ensuring privacy is preserved and personal data is protected.

An illustration of this is market research commissioned by the public sector.

At first sight much of this is already Open Data. However it is not fully Open Data. Innovative applications may require data not asked as part of the survey to be appended to the responses. Typically this involves matching data for individual respondents and to achieve this it is necessary to create processes that achieve this match yet also protect the personal data.

CACI's experience is that in some case the public sector will make big efforts to find constructive ways within the appropriate data protection processes. (For example ONS proved extremely helpful in relation to access to the Expenditure and Food Survey.) In other cases we have found that despite a government department that has commissioned research being happy to allow access there can sometimes be little cooperation from the private sector research organisation that carried out the survey and is responsible for the personal data of the respondents.

We have found the greatest good comes where there is a trusting relationship between the parties, each appreciating the constraints of each other's data governance policies. A coercive framework of sanctions is unlikely to prove helpful in such cases.

4. What other sectors would benefit from having a dedicated Sector Transparency Board?

Areas relevant to us might include sectors such as Land, Property and Social Statistics.

Meaningful Open Data (Page 31)

1. How should public services make use of data inventories? What is the optimal way to develop and operate this?

If private sector companies were obliged to compile and publish perfect inventories of every form of data they held it would be considered damagingly onerous.

The delay required to compile perfect data inventories will damage the Government's open data approach. Therefore we suggest a more pragmatic approach. Many businesses and academics use or know of data sets held within the public sector. Many public sector employees know of databases that they

use. Ask these people to submit their list of databases to a common inventory and use this as a starting point. There will inevitably be gaps, and some gaps may need filling to achieve greater accountability, but these can be addressed over time.

We believe the benefits of creating some open data quickly outweigh the disadvantages of an incomplete data inventory.

2. How should data be prioritised for inclusion in an inventory? How is value to be established?

In the short term a good criteria is availability. Focus on existing data sets that can be published in raw form without the need for government to spend time and resources processing or interpreting the data. The process of prioritisation is likely to delay publication of data.

There seems no good method by which the public sector can establish the value of a data set from which the private sector may generate value-added products or services.

For example, we think the current approach whereby the Land Registry decide whether or not a particular use of data is beneficial is inefficient. The Land Registry should require the user to conform to all legislation, and may reasonably prohibit activities that might harm its future collection of the data. The (private sector) user should determine whether there is value in the use of the data.

3. In what areas would you expect government to collect and publish data routinely?

We suggest that there may be merit in considering different categories of data. For example, a data set such as the National Address Register has many implications for the stimulation of economic growth but seems to have less application in the context of increased accountability.

Trying to cover all categories and uses of data within a single web site such as data.gov.uk runs the risk of hiding some datasets behind the demand for others. (As an organisation seeking data for the purpose of building value added products and services we find data.gov.uk to be of little use. The site may include information useful for our purpose but it is obscured by the pushing of data intended to increase accountability.)

We accept that other private sector organisations may make use of data apparently better suited to accountability.

We would like as a minimum to see the government publish the following (including publishing in processed and enhanced formats)

- Socio-demographic characteristics for Census Output Areas (This includes all data sets being considered by ONS as part of their Beyond 2011 project.)

- Market research survey microdata covering citizen / consumer activity relevant to government policies or statutory requirements of local authorities or other bodies
- Anonymised *local* information (preferably at Census Output Area level whenever this can be consistent with protection of personal information) derived from all research surveys and registers of data collected by government departments or agencies. This would include but not be limited to; Land Registry, VOA, DVLA, HMRC, DWP, HESA, The Rent Service, DfES, NROSH, DfE, BIS, Home Office

4. What data is collected "unnecessarily"? How should these datasets be identified? Should collection be stopped?

We suggest that at this stage it is difficult to identify whether some data is collected unnecessarily. Data considered by some to be of no value may be valuable to others *who are not yet aware of the existence of the data.*

To date the data may not have been made available to users, or not with the full flexibility that we suggest in our other responses. As a result (private sector) users may not yet have had the opportunity to evaluate whether the data is of value, both in itself and as input to value-added products and services.

A rush to stop collection of data risks losing potentially valuable data. We suggest the priority is improving access to data. Identification of unnecessary data collection is best done once the data has been 'open' for some time. "Unnecessary" may then be identified by a) readily accessible data that is accessed infrequently and where b) those infrequent occasions are seen to be of little value.

5. Should the data that government releases always be of high quality? How do we define quality? To what extent should public service providers "polish" the data they publish, if at all?

For the purposes of stimulating economic growth raw and poor quality data is always preferable to no data at all.

Moreover our experience generating value added products and services leads us to conclude that raw data can be of greater value than "polished" data. This is quite normal, most of the private sector data used by CACI can be considered incomplete in some way. We note that in presenting their Beyond 2011 project ONS acknowledge that the private sector is "good at using messy data" and so may provide assistance to the project.

Having published the raw data there may also be a case for public service providers to invest resources to "polish" the data. This may be of less benefit in stimulating economic growth, rather it may provide gains in accountability, efficiency, and citizen engagement.

It is important to realise that the process of polishing implies that the public service provider knows and understands how users wish to use the data. It seems

likely that polishing the data will enhance the data for some users and make it less useful for others.

An example of this is the Land Registry 'Price Paid' dataset. This dataset has been well-designed for the purposes of earliest users of the data. Unfortunately admirable willingness to respond to the needs of those early users, combined with a policy of fairness insisting the same dataset must be provided to all results in a data set that omits some records and data fields and so is less useful than it might be to other users. It is frustrating to know that data we could constructively use exists but the polishing done for others means we cannot have it.

Government sets the example (Page 33)

1. How should government approach the release of existing data for policy and research purposes: should this be held in a central portal or held on departmental portals?

Currently CACI are able to access much open public sector data through a mix of departmental portals and some (partially) central portals. There is no clear advantage to either approach. They key thing is the identification of contact points to allow the user to enquire about the data and to make requests for beneficial (and secure) appending of other data to a data set.

We note that we have found the 'central portal' of the ESRC Data Archive to be beneficial. This is primarily due to the willingness of the data archive staff to facilitate contact with the organisations placing data into the archive. The same opportunity to identify contact points is true of data sets available through the ONS Neighbourhood Statistics web site.

Many years' experience using open public sector data has taught us that most is achieved when people talk to people. Relying exclusively on automated technological approaches will miss many opportunities.

2. What factors should inform prioritisation of datasets for publication, at national, local or sector level?

As mentioned previously, we think that in the short term the key factor criteria is availability. The focus should be on existing data sets that can be published in raw form without the need for government to spend time and resources processing or interpreting the data.

The private sector will best stimulate the economy if data is published in as much detail as possible, within disclosure constraints. For example; the value of much socio-economic data is destroyed when it is only available at high levels (such as Region or Local Authority). Publication of data at the lowest possible geographical level (such as statistics at Output area level) will produce a significant benefits.

3. *Which is more important: for government to prioritise publishing a broader set of data, or existing data at a more detailed level?*

   The value of currently available public sector data is significantly diminished by a tendency to publish the data at unnecessarily crude geographic detail. In many cases little or no effort would be required to publish at greater detail. Specifically the public sector appears to believe the Lower Level Super Output Area (LSOA) provides sufficient detail – yet CACI find clients in the public sector eager to use data for smaller areas.

   We believe a 'quick win' would be for the government to prioritise an investigation of the reasons for less detailed geography becoming the de-facto standard and if appropriate implement guidelines to change this where possible.

   In this regard we note that the census publishes detailed information at detailed Output Area level. ONS have processes to do so while taking care to protect individual privacy. These might offer the way to providing more geographically detailed data, if there is the will to do so.

Innovation with Open Data (Page 36)

1. *Is there a role for government to stimulate innovation in the use of Open Data? If so, what is the best way to achieve this?*

   In our view an effective way to stimulate economic growth is to encourage innovation and increased data use is be ensuring raw data is both free and easily accessible. We believe a competitive environment where private sector developers produce market driven products and value-added services will be both innovative and efficient. Most importantly we believe it is the best way of producing services appropriate to the needs of end-users.

   This efficiency and innovation will be enhanced by access to the raw data (however 'messy') in basic text file formats, a secure and appropriate approach to data confidentiality (as discussed above), sufficient metadata, and by avoiding complex licensing and excessive charges.