IBM's Response to HM Government Public Consultation: Making Open Data Real

1. Executive Summary

IBM welcomes the opportunity to submit this response. In this document we respond to the questions raised for consultation, and offer some additional insights which we hope will be of assistance.

In our view, the top challenges, any of which could derail the efforts and success of the Open Data initiative, are:

- the charging and funding model for making data available
  - including hidden costs and making data affordable
- security and privacy considerations

Practical solutions exist, with varied pros and cons, and we suggest some approaches.

This consultation has a stated "presumption that data is open by default". There is some tension between this and the approach put forward in the Government's Consultation on Data Policy for a Public Data Corporation, which discusses a charging and funding model which in our view is likely to affect the " open by default" aspiration. As part of the overall planning for Open Data, these different approaches to policy will need harmonising.

2. Responses to Consultation Questions

Glossary	
1	**Do the definitions of the key terms go far enough or too far?** The definitions are a good starting point, but on their own are probably too broad to fully support the formation of a workable set of principles, guidelines and legislation. For instance, resolving the tension between data made open in the public interest and data kept private to sustain competitive delivery organisations will require more granular definitions. Ideally, we need to be able to define such things as data taxonomy, security classification, ownership, quality etc. This would position the glossary to be able to describe most aspects of data and enable it to be used for current and hopefully future needs.
2	**Where a decision is being taken about whether to make a dataset open, what tests should be applied?** The first test should be to determine if the dataset is useful, or likely to be useful in the future, and if the cost of publishing it is likely to give value for money. Forecasting can be difficult and inaccurate, and a principle of the Open Data initiative is that innovation and unexpected value will be created if data is opened up, so the default assumption will 'yes' ('Open by default'). Other tests should consider reasons which might prohibit release, the key ones being privacy and security. In addition to privacy of the individual citizen, commercial and competitive privacy for enterprises needs to be taken in to account. For example, private or Third Sector organisations who will be competing in the envisaged market for Open Public Services will need some assurance that data concerning their delivery of service will not be made open where it might damage their competitive position.
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?** The answer to this depends in part on who decides what is value for money. In many cases the money will be the provider's and the value will be the requester's. This is not usually the recipe for a successful business case. It could be argued that no system or process in place before Open Data regulation will have been designed to support Open Data, and that any Open Data requirement will constitute a new cost. One approach would be to require that all new systems and processes are designed and costed with the ability to provide Open Data. For pre-existing systems, some areas of Open Data could be the subject of legislation requiring that changes are retro-fitted (as, for example, changes in mortgage regulations in the 2000's required mortgage providers to re-engineer their systems). In other cases, where the cost of providing Open Data from pre-existing systems and processes exceeds a given threshold, the presumption could be for the requester to pay.
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?**

	Please see our response to question number 2.
5	**What would be appropriate mechanisms to encourage or ensure publication of data by public service providers?** • Legislation, regulatory and contractual approaches • Paying for the data, for example by permitting providers to charge users • Societal pressure that publication is the expected norm, and part of service providers' corporate and social responsibilities • Incentives (e.g. grant support) for fostering entrepreneurial activity around the use / exploitation of open data (ultimately driving knowledge worker job creation.)
Policy Challenge	
1	**How would we establish a stronger presumption in favour of publication than that which currently exists?** The approach laid out in the Consultation paper looks very good. The difficulties are going to be in getting the approach right to cost and open data, versus competitive and commercial considerations, The latter will depend heavily on being able to clearly define the competitive context (including open data) that providers are required to operate in when tendering.
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?** We would support, as part of a wider picture, an independent body such as the Information Commissioner, which would include right of appeal to challenge cost-based refusals to make data open. Consideration should also be given to a regular "Information Audit" that reviews (or samples) all cost assessments, whether or not they are challenged.
3	**Are existing safeguards to protect personal data and privacy measures adequate to regulate the Open Data agenda?** There are unavoidable risks and issues associated with releasing Open Data, whether the existing safeguards are adequate depends on the level of risk Government is prepared to take in order to realise the benefits of Open Data. Please see Section 3 for more in depth discussion of privacy and security issues, for example: • A key risk is data aggregation, where combining of datasets can inadvertently reveal information which can be linked back to an individual citizen. • An important technique to address privacy issues is anonymisation of data, though it is not foolproof. A number of government department already deploy this in ways which preserve data's essential characteristics but prevent linkage to personal information, for example. Making information available in this fashion still allows trend and pattern analysis. **Scope of 'public' data :** The consideration of privacy issues must include discussion on the scope of what is 'public'. Public interest is subject to economic value arguments, and may result in citizens or enterprises seeking data that goes well beyond what is usually thought of as public data. For example: suppliers' and contractors' performance against service level agreements, activity of highly regulated sectors such as financial services or telecommunications, and government franchises such as transport, often have static and dynamic data that may need to be considered as part of this open data debate. The data might need to be used and reflected in government contracts, franchises and auctioning of government controlled public goods. Yet this data may be highly commercially or competitively sensitive, posing

	suppliers an issue with the impact of its release.
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?** The resource implications might be significant and prohibitive. Making requirements explicit in tendering processes will partially address this. Please see Section 3.1 for more discussion of this issue.
5	**How will we ensure that Open Data standards are embedded in new ICT contracts?** • In the first instance, the set of Standards needs to be communicated to all stakeholders. This is more than simply posting them on a website; it will require a comprehensive communication plan to be worked up and delivered, probably involving multiple communications via multiple media channels, to all affected parties. • The set of standards should be specified (i.e. mandated) in tenders. Suppliers should be asked to indicate those standards that their solution does not meet (with rationale, and potential mitigations). • Use of the standards should be tested for at appropriate points in contract negotiation and factored in to contract award. • A system of governance will be needed to track standards compliance from contract in to delivery. • An escalation process will be needed to consider of exceptions to the standards – either variants or requests for waiver - , so that an appropriate decision can be made which takes in to account business conditions. Such exceptions should be fed back in to a vitality process for keeping the standards up to date, • It will be necessary to 'prove' this governance system i.e. demonstrate it in action, with tenders being rejected, because they do not meet the standards. • Suppliers are more likely to use and comply with the standards it they are involved in selecting and updating them, however it will be necessary to centralise the creation of a set of standards – there is not enough time or skilled local resources to do this well across the UK with any approach. We applaud the intent behind Government's crowd-sourcing standards survey earlier this year, though expect this might not necessarily be delivering either a complete set of standards, or made it easy to give appropriate focus to those standards which will have significant impact compared to those with minority impact. In addition, the survey might have erroneously raised expectations amongst respondents that their responses will be acted upon and satisfied – this is 'mission impossible' where conflicting responses have been received. Please see Section 3 for more discussion of this issue.
Setting Open Data Standards	
1	**What is the best way to achieve compliance on high and common standards to allow usability and interoperability?** • Ensure the standards are thoroughly communicated to all stakeholders. • Use of the standards should be tested for at appropriate points in service design, build, deployment and maintenance. A gated project lifecycle (such as those used in delivering most ICT solutions) typically offers a number of opportunities for this, and risk-based options (e.g. waivers, or 'light touch governance' for low risk contracts or solutions) can be considered to minimise cost. • As for standards considerations in contracting, a system of governance will be needed to track standards compliance in delivery, and also provide an escalation process for exceptions to the standards, so that an appropriate decision can be made which takes in to account business conditions, with the exceptions fed back in to a vitality process for keeping the standards up to date.

	• It will be necessary to 'prove' this governance system i.e. demonstrate it in action, with projects being stopped, because they do not meet the standards. This potentially costly action probably only needs to be demonstrated a few times, early on, for the governance to be taken seriously and the standards acknowledged and adhered to. • Annex 2 of the Consultation paper documents a set of Public Sector Data Principles. These are a valuable complement to formal standards. Experience suggests that when faced with a set of rules (such as standards) there will always be attempts to work around them by stakeholders who find the rules do not entirely suit their interests. Issuing the Principles with an instruction that they are to be followed not only in the letter but also the spirit, is a useful adjunct to (but does not replace) formal governance against a set of standards. **Five Star Rating for Open Data** Tim Berners Lee's idea might help drive standards compliance; it could be possible to set expectations for a 'star rating' for certain classes or domains of data, based on an analysis of the availability of relevant standards for that type of data.
2	**Is there a role for government to establish consistent standards for collecting user experience across public services?** Yes, there is a role for Government to establish consistent standards since without this consistency it will be very difficult or impossible for service users to make comparisons and choose between alternative services. The existing framework is quite fragmented, and if a consistent and successful Open Data approach is to be delivered, a single body will need to be responsible. But care should be taken that this done in a way that would not inhibit 3rd party innovation in this process.
3	**Should we consider a scheme for accreditation of information intermediaries, and if so how might that best work?** We assume the term "information intermediary" means, loosely, organisations which manipulate Open Data. We support the notion that they are accredited, since it will contribute to the drive for data quality, and will increase user confidence in using Open Data.
Corporate	**and Personal Responsibility**
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**. In the first instance, the presumption to publish Open Data should be built in to the day to day business processes of service providers. But beyond that, personal accountability is key. For each service providing organisation, there must be an identified responsible individual, with appropriate authority, possibly at Board level, to ensure the commitment is honoured. Privacy and security considerations need to be factored in, but their criteria should be transparent, and instances when security and privacy are used a reason to refuse Open Data publication audited and challenged on a regular basis.
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?** An executive Board member, with personal responsibility for Open Data, would be in a position to ensure that business process support the right to data, and put in place audits and other governance approaches to ensure it is delivered. Data protection and privacy concerns inevitably precipitate a risk-averse "don't publish unless you have to" view of data. In order for proper consideration of the drivers for the right to data to be considered against this, it is best that the responsibilities are held in

	separate roles by separate people. The debate can then be held transparently.
3	**Would we need to have a sanctions framework to enforce a right to data?** Yes; some mechanism is required to monetise the enforcement of Open Data
4	**What other sectors would benefit from having a dedicated Sector Transparency Board?** Potentially there is a very wide set of sectors that would benefit from a Sector Transparency Board, it includes any that deliver services of economic or social value in the UK. In the first instance, health, judicial and education services might be at the top of the list, and also some Defence sector divisions.
Meaningful Open Data	
1	**How should public services make use of data inventories? What is the optimal way to develop and operate this?** It will probably be impossible to create a perfect navigation scheme for any person or organisation to find the open data of interest to them. An attempt to do so would constitute an extensive Enterprise Information Architecture exercise. Some such schema (a data directory) *should* be created as a starting point and default; but techniques for navigating open data should themselves be the focus of community innovation, either by simple approaches such as allowing data consumers to "tag" the data according to its meaning to them; through to more involved exercises such as third parties implementing their own navigation portals for open data - some of which could even charge for use.
2	**How should data be prioritised for inclusion in an inventory? How is value to be established?** There is no single answer that can be applied generally. • For some data, a clear business case could be constructed to justify investment in its publication. This would be based on metrics associated with the value of the data e.g. how many times it is used, impact it has when used, mapped against the difficulty and cost of making it available. • Other data may be cheap enough to make open that it is obvious to do so. • There may need to be an additional process for stakeholders to make non-quantified cases for the release of additional data, followed by a trial, followed by consideration of a business case for full publication.
3	**In what areas would you expect government to collect and publish data routinely?** The potential list is very long, but includes publically delivered services, and publically owned assets. As well as the large services and assets such as health, schools, etc, there is a huge number of smaller assets and services e.g. park benches, public toilets, dropped kerbs, grit bins and so on. As well as cost, service and performance levels, usability information such as addresses and opening times should be published routinely. Census data is already published, and information held in the National Archives and other publically funded repositories should be digitised and made open so far as possible – the task is immense and will be subject to cost/benefit prioritisation. But technology advances, as well as approaches which make use of volunteer effort (similar to crowd sourcing) are advancing rapidly to make this increasingly cheaper and more viable.
4	**What data is collected 'unnecessarily'? How should these datasets be identified? Should collection be stopped?** It is extremely difficult to identify "unnecessary" data in this context - the innovation around

	Open Data that the government is seeking is the very process of entrepreneurially finding or creating value where others have not. A more useful approach would be for organisations to self-select data to stop collecting driven simply by what they can and cannot afford to do.
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?** It is more important to let users know the level of confidence they can have in data rather than go for 100% accuracy (which might also incur legal liability issues)**.** Even incorrect data can have value, though it should be appropriately flagged as being suspect or unconfirmed). Any polishing should only be performed to ensure ease and consistency of access.
Government sets the example	
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?** Both central and department portals will be needed. Some data searches will clearly be most obviously directed at a departmental portal. But the intent behind other searches will have no relevance to existing government silos, and a federated approach will be needed. In ICT terms, some of the datasets will be segmented and distributed across departmental systems according to pattern of use.
2	**What factors should inform prioritisation of datasets for publication, at national, local or sector level?** The key factors are value versus cost, but it may be hard to get this calculation right, since any assessment of value is likely to omit the value that can be created by unforeseen uses to which data can be put. A reasonable default position is that after the most useful data has been published, anything which can be published cheaply should be published. Service performance data would seem to be another candidate for early publication since demand appears to be high – though probably research will indicate which services to prioritise.
3	**Which is more important: for government to prioritise publishing a broader set of data, or existing data at a more detailed level?** It is most important to publish the most useful data first. Beyond that, this question is very sensitive to context, and the value in the data that potential users might perceive. A meaningful answer probably has to be worked out on a case by case basis.
Innovation with Open Data	
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?** 1. To stimulate innovation, Government needs to show and publicise that Open data can change things, enabling citizens to take control of their own destinies, making more informed choices that change their lives, enabling them able to spend money in the way they feel is best for them.

	2. As well as making sure that Open Data is high quality, accessible and understandable, Government could encourage usability features such as a publish and subscribe capabilities, which would result in Open Data becoming part of everyday life e.g. citizens checking on performance of local services is likely to drive opinion and proactive service improvements. 3. For enterprises to identify, access and manipulate data to create a commercially viable innovation, some enabling platform / environment needs to exist which they may not be able to fund in advance. 'Build it and they will come' initiatives have so far had limited success – the business case is difficult for the private sector, and no easier for public or Third Sector. So there is also a role for Government to stimulate innovation in providing this platform or environment.

3. Additional Insights

3.1 Charging and funding for making data available

Should a publically funded body be able to charge for making its data open, and pursue private investment? The Open Data and "Public Data Corporation" proposals are not solely an "Open Data" proposal. They are a combination of 1) improving access to Open Data and 2) recognising the business models of public agencies whose primary remit is to create and disseminate accurate information (as distinct from agencies which produce data as a natural 'side effect' of their primary mission).

The data products and services of publicly funded agencies could be treated in one of three ways:

- data which can be sold (typically by licensing)
- data which can not be sold – no market currently exists
- and probably, data which is in the grey area between the two categories above.

3.1.1 Data which can be sold

The UK has a number of public agencies which incur cost usually by undertaking extensive survey, analysis and data. Examples include Ordnance Survey, the UK Hydrographic Office, the Met Office, the ONS which deliver data of huge economic value, and which can be sold for a reasonable rate of return or at cost. Examples of that data include:

- Maps and nautical charts
- Addresses and postcodes
- Location of surveyed buildings, roads and other assets of interest
- Weather station readings
- Demographic data (the census)

In many cases, the agencies are already charging for these datasets, and in so doing are open to the accusation that they are government monopolies, involved in price-fixing. Both cost and reasonable rate of return are difficult to derive, but must be made transparent. There has been some public controversy over this aspect of the proposed Public Data Corporation. It is not hard to see why as it overlaps with the interests of some organisations in the private sector who seek to generate revenue by processing information, and which do not receive public support.

3.1.2 Data which can not be sold

Other datasets do not currently have, and it is not known if they will ever have, commercial value. An example might be survey data related to very remote parts of the country such as uninhabited islands with no population or known resources.

3.1.3 Data sets which are in the grey area between the two categories

Examples might be:

- survey data which would support an activity which is not currently commercially attractive and for which no business case exists, but might be deemed important or necessary in the future, e.g. supporting the extension of broadband to remote areas
- historical data preserved to support future knowledge workers in the creatives industry such as film, TV, music

3.1.4 What can we learn from the experience in other countries?

Monitoring the experience of the impact of Openness on cost and dynamics of public data in other nations could be very helpful.

A case study might be the US equivalent of the Ordnance Survey, which does not charge for data. A large number of commercial cartographers and mapping services in the US now exist, which might indicate that the free and open access to the data has either

- stimulated a healthy and innovative market, producing products and services of wide benefit

OR

- created cost pressures (perhaps due to lack of the revenue stream that selling the data would have provided), which have driven down the quality of the publically available data to the point where it is of reduced value, and commercial organisations have been able to easily step in. If true, this wastes public money, and private sector cartographers may have no obligation to provide complete coverage, keep maps up to date, or use standardised mapping notations which would support data exchange and interoperability.

Conclusion:

A principle of the Open Data movement is that unexpected value will be created if data is opened up. The corollary is that it is difficult or impossible to know in advance the value of data that might emerge from such exploration. If the cost of creating it were not publically funded, who is to say the private sector would ever invest to create it? IBM believes some level of public funding is needed to allow for collection of datasets which cannot be sold, or which are in the grey area, and it is key is to develop criteria which will support decisions of which datasets to fund, and which to let go or leave to market forces.

3.2 Hidden costs in making data available

Some data will be very difficult for non-specialists to understand and use properly, and cost will need to be incurred to address this. For example the London Data Portal makes available data that provides an overview of the health of the labour market by borough. The data includes an "employment rate" quoted as a percentage. Is this the absolute number of people employed as a percentage of the overall population? Or only those of working age? Or do some other criteria apply? This is a simplistic example, but it illustrates an important point - Open Data will only be useful if it is accompanied by clear explanations - which will increase the cost of providing it. A data release process could be designed that would deliver this, perhaps supported by a ‘customer service’ element – a further element of cost.

The report states "Fundamentally, the right to continued access to a dataset, once released, does not exist.". If a right to continued access is created, how frequently can that right be exercised? With what latency should data be available? In extremis this discussion reaches real-time (or near real-time) data which can be extremely difficult and expensive to provide, but is the ‘fuel’ for many valuable services e.g. real-time reporting on transport delays with alternative route planning, dynamic management of utility (power, telecommunications bandwidth etc) supply and demand .

In many cases, investments will need to be made to make data open before anyone even realises where the eventual value will be realised, or what it will be. Some potentially costly infrastructures - such as a really effective data catalogue - will need to be created to allow consumers to find the data

they need. All of this implies that progress will be halting and slow unless some level of appropriately governed central investment is available.

3.3 Addressing Affordability - making data available over time

Comments in the Consultation paper (e.g. in section 8.6) recognise that the costs involved in making data openly available are closely associated with the processes and ICT systems that manipulate that the data; and that the economics for making data open will sometimes only work as systems are upgraded or replaced.

We agree with this assessment. Open Data is not going to be free to provide, and its provision in some cases will need to be seen as a journey throughout the course of which central government, public sector agencies, suppliers and information consumers all make investments. This type of transition is not uncommon in the management and development of large ICT systems, and there are established techniques (e.g. Enterprise Architecture approaches) for delivering good outcomes, with changes delivered over time. Nevertheless, a timescale, priority and business case is needed to make sure the objective remains 'on the agenda', and is not simply progressively deferred in favour of cost savings as each individual upgrade or replacement project is commissioned.

3.4 Security and privacy considerations

3.4.1 Data aggregation

It is relatively straightforward to assess the risk in making available individual instances of data, whether they are historical records and reports or live performance feeds. However there is risk that insight could be derived when multiple sources of data which are made available are combined. These risks might include:

- inadvertent exposure of information which can be linked back to an individual's private circumstances
- inadvertent exposure of details of government operations, locations
- other unforeseen situations.

This risk is compounded when considering not just what information has previously been made available, but what might be made available in the future (including, perhaps as a special example, open source intelligence data held in social media tools) . So, the decision making process to allow access to information should consider the risk of data aggregation arising from successful future Freedom of Information Act requests

IBM's experience is that this risk is best managed by a single register of available data, managed by a central authority whose remit includes due diligence of candidates for open data – potentially this role could be taken by the proposed Information Commissioner.

3.4.2 Basic security

The systems which provide access to open data must be implemented and operated according to IT security best practices. Wider access to government data widens access to electronic attack. From a security perspective, the fundamental approaches to address this are that :

- systems should be designed to be protected from distributed denial of service attacks
- they should be separate from systems which sustain departmental operations and hold personal and other sensitive data
- they should be physically protected
- activities of those who undertake publication of information should be audited.

3.4.3 Data protection through anonymisation

Government may wish to publish information which has great economic or social benefits, but contains sensitive (usually personal) information. One well established technique already employed by a number of Government agencies is the anonymisation of part or all of the data set , in ways which preserve its essential characteristics but prevent linkage to personal information, for example. Making information available in this fashion allows trend and pattern analysis.

3.4.4 Near real time data

Performance and status information is likely to be most valuable on a near real time basis – for example traffic and transport delays, extreme weather reports and so on. Such information may present undue risk in both envisaged and unexpected circumstances. For example, in a raised Threat situation, it might be decided that a data feed could be exploited by an adversary. Alternatively an emergency may occur and malicious opportunism becomes a greater risk.

In both cases a response such as suspending the data feed, or perhaps providing it on a degraded basis which is not obviously apparent to consumers of the feed, might offer solutions.

3.4.5 Revoking access

Licensing policy and approach should allow for the license to access live data by a consumer to be revocable.