NHS Information Centre for health and social care
Responses to the Consultations on:

Making Open Data Real

Data Policy for the Public Data Corporation

27th October 2011

Summary

The NHS Information Centre (NHS IC) welcomes the opportunity to contribute to these consultations. We recognise that Government is committed to maximising the utility of public data for a number of purposes, including the stimulation of innovation and growth. This is predicated on the assumption that there will be some creative tensions as well as some unintended consequences. We welcome the opportunities created by the Open Data agenda.

The collection and provision of data is the core business for the NHS IC. We therefore have a material interest in the future direction for Open Data. We have a major role in unlocking the potential for making better use of information, especially in regard to the use of data for secondary purposes such as policy development, planning and commissioning, service improvement, research, public accountability and transparency.

We have combined our responses into a single document on the basis that some of the underpinning principles ought to apply to both consultations. Before getting to the specific questions raised in the consultation, there are some general comments we would wish to make:

- The establishment of the Public Data Corporation, and the Open Data agenda, are both means to an end – and that is to improve our collective ability to generate and use data for a range of purposes, to enrich society. This is a major opportunity to champion the value and utility of data and intelligence, and has implications for data which should apply prospectively and retrospectively. In focussing on making data more accessible, we are inevitably concentrating on data which is already collected and available for publication. The data currently collected is a result of many years' policy decisions. In the context of health and social care it is subject to regular review. We know that it is often difficult to stop a data collection, even when its utility is questionable. We know also that there are gaps in terms of data coverage at national level. Open Data is an opportunity for a fresh look at the way we use data.

- The bigger challenge to encourage innovation and use of the data is a very complex one, which goes beyond the proliferation of raw data. People and organisations will able to make better use of data if they have confidence in the data. The full benefits of Open Data will only materialise when people, businesses, researchers and other organisations understand the data, and can use it and interpret with confidence that it is fit for their specific purpose.

- To facilitate effective use, it is important that any release of data is partnered with sufficient contextual meta-data to adequately describe the contents, the issues and pitfalls associated with that data. Sometimes post-release user / data support will also be required to handle queries about the release materials – there is a cost to the releasing organisation of such provision, and it is extremely difficult to predict what that is likely to be, until Open Data is a more settled concept.

- The routine publication of data in raw formats, of itself, will not enhance a citizen's access to a range of data or ability to analyse the data to support Choice. It will take

time for the information market place to develop in ways that benefit the public. Commercial, third sector and other information intermediaries are significant users of our data and services. The NHS IC is already working collaboratively with these organisations to support the diverse, strong and responsive information market that this agenda requires.

- The NHS IC has a national role in publishing and managing Official and National Statistics. A statistical approach to managing public data, as outlined in the UK Statistics Code of Practice, can complement the routine publication of Open Data by:

    - Ensuring that data are collected, published and presented in a standardised way, that is accessible and easy to use;
    - Providing information ("metadata") about the data to help people know what is (and is not) a fair and reliable use of the data;
    - Ensure that the data are recognised as having integrity so that there is confidence in the applications and conclusions drawn from the data.

- Technologies are changing the way we think of "data collections". The consultation document occasionally positions these as being central to transparency. However, we are now able to extract data "on demand" rather than use regular, routine collections. If national organisations no longer collect the data, then the focus of responsibility shifts to local organisations. This will raise further issues about resourcing this work. It also puts a reliance on local host systems and at times transactional systems – which will bring more information governance issues to the fore.

- The Open Data agenda is sometimes ambiguous in the way it deals with stakeholder interests. Sometimes public interests are given greatest importance; sometimes a more agnostic approach is adopted. This is an issue which needs greater clarity.

- We welcome the recent report[1] published by the Cabinet Office about the need to balance the interests of privacy and transparency. This is an important contribution to the national debate, and it would be helpful to have an early indication from Government as to how it expects these issues to be addressed.

- The role and functions of the PDC need to be understood in the wider context of the Open Data agenda. The "direction of travel" for Open Data affects all organisations delivering public services, and uses a broad definition of public data. Whilst current thinking on the PDC is focussing on those organisations which operate as trading funds (OS, Land Registry and the Met Office), it is clear that the establishment of a PDC and the policies surrounding it will have implications for the wider Open Data agenda, and therefore for the NHS IC also. These include:

    - The scope and concept of public data, and how it fits with the PDC;
    - Concept of the PDC and the organisational form it takes; - is it an entity or a federation of organisations? A delivery mechanism or a commissioning mechanism?
    - From this, what are the implications for organisations and services not included in the PDC? Where are there matters of principle, or specific issues where there is a

[1] http://www.cabinetoffice.gov.uk/resource-library/independent-transparency-and-privacy-review

need for consistency across public services/public data (eg implications for intellectual property, the Open Government licence).

The NHS IC looks forward to contributing to the debate and the transformation involved in the Open Data agenda.

1. Our response to the consultation – "Making Open Data Real"

Glossary

1. **Do the definitions of the key terms go far enough or too far?**

We know from our experience specifically in the health and care arena that terminology is important. Terms such as "data", "data sets", "records" (including medical records) often mean different things to different people, It is therefore important that there are clear, unambiguous definitions as to what is meant in regard to the policy and principles involved, as well as the detail. These should be applied and adopted across all organisations and services. Otherwise we risk a non-alignment of expectations and delivery.

The list of items in the glossary do not go far enough to bring sufficient clarity and understanding to the Open Data agenda. We appreciate that there is a need to balance the degree of specialist understanding associated with these terms, with the need to communicate effectively to different audiences, including the public. This will not be easy to achieve, however, as the recent Cabinet Office report "Transparent Government, not Transparent Citizens" makes clear.

Therefore there is a need to keep the glossary to a manageable and meaningful level, but it would be useful if its contents went further than looking at terminology relevant to raw data. It should also include terms such as "standards" and "inventory". It would also be prudent to avoid any circular uses of the word "data" to define terms such as "data" or "dataset".

Although this is a pan-Govt initiative, it would be reasonable to expect that it would use established expertise and practice. A good example would be the work of the Information Standards Board for Health and Social Care, which has established a common set of definitions and standards that are in use across the NHS.

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

The tests need to be clearly understood, and applied consistently. A checklist approach would be helpful so that there is clarity as to the considerations that need to be addressed, and by whom. Factors which need to be considered will include:

- What are the data items to be made available?
- What metadata needs to be made available?

- What is the data source? Is it routinely available? If it is not available, what needs to be done to make it available?

- What processing is required to make the data available? How much processing might the users be expected to do?

- In what format will the data be made available? How big is the data file?

- How frequently will the data be updated? Will it reflect current update processes, which are predominantly annually and quarterly? Or would it be possible to provide monthly updates?

- Is the data available at the right level of granularity (by commissioner and provider). Can the data be reprofiled to reflect organisational configurations?

- Are there any access issues or constraints? How will they be addressed?

- Does the data include any patient-identifiable data? If so, can these items be removed or protected? If not, the data should not be published.

- Does the data require the use of statistical disclosure controls, for example to manage the occurrences of small numbers? What controls will be necessary?

- What additional guidance or contextual data is required, if any?

It is not intended to make assumptions about relative value or utility of data for different users – there is a recognition that data can be used for different purposes, and the primary objective is to encourage this to happen. Therefore it is likely that the main factor in deciding to publish will be the cost of doing so. Any attempt to create a hierarchy or framework of interests based on perceptions of utility or value would risk undermining the strategic intent behind “Open Data”.

It is important to repeat the commitment not to compromise public trust. “Open Data“ will not publish anything from which people can be identified. Whilst it is relatively easy to assess each candidate data set for its suitability for publication, it is necessary to recognise that jigsaw reidentification techniques can be used to combine different datasets which could result in an ability to identify or infer identities of individuals. This represents a separation of the risks associated with publication and the risks associated with the use to which the data is put.

Moreover, this will not be static – datasets will be released at different times; releasing one dataset in 2011 may itself be straightforward, but a different dataset published in the future may open up subsequent opportunities for jigsaw reidentification. Therefore, there is a need to consider whether decisions to publish datasets should be taken in isolation, or whether there is a need for a central “oversight” role.

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?**

There are three categories of cost – cost to collect, cost to publish; cost to use. All incurred by different people/organisations. And in many cases these will not be the people/organisations deriving the benefit from the data.

It should be relatively easy, though, to establish some criteria for charging based on the amount of additional processing required to make the data fit for publication over and above that needed for its original purpose.

The use to which the data will be put will determine the extent to which the user may be prepared to pay. A member of the public may be less keen to pay for data than a commercial or research organisation. This has implications for both the information marketplace and the rate of take-up and use of the data. Government may wish to consider whether there is a trade-off in terms of charging for and use of the data.

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 do not offer a direct response to this question, but offer three comments which are relevant to this:

- The definition of "Open Data" will be key to this;
- Whatever decision is taken, it needs to be communicated clearly and applied consistently;
- There needs to be some consistency for national and local organisations;
- It is important that the threshold does not bring unintended consequences, eg for smaller organisations (third sector, or social enterprises) which may undermine their business model. This decision cannot be considered in isolation of a decision about charging. Some organisations or services may incur more costs in processing and publishing the data.

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

If the principles and definitions are clear, and the arrangements for covering costs were fair, there would be no need for mechanisms.

An enhanced right to data

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

The most important considerations relate to clarity of principles, definitions and requirements, so that there is a common understanding across data providers and consumers.

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?**

   It seems reasonable that the Information Commissioner's Office should have enhanced powers to fulfil this role. This is preferable to the establishment of a new body, as the ICO already has a public profile in this space.

   One of the difficulties will involve the arrangements which apply to different data published by different organisations, especially where there are apparent inconsistencies. Therefore, it is just as important that there is a broad understanding of the roles and responsibilities of other organisations, such as The National Archives.

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

   As with other issues relating to Open Data", this is not a static issue. Different people and organisations will inevitably have different interests and perspectives on this. The recent report published by the Cabinet Office provides a helpful analysis of current policy thinking, and this makes it clear that there are still uncertainties around public expectations and behaviours.

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?**

   Our comments in response to question 2, about deciding to make datasets available, and question 4, about the organisational scope, are relevant here.

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

   We offer no response to this, but again comment that this will not be a static issue – it will change over time as technology changes, and there is more clarity about the demand for and use of the data.

Setting Open Data standards

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

   Engagement with stakeholders in the development of the standards is more likely to guarantee compliance. It is important that the timelag for developing, agreeing and implementing standards is factored into the Open Data agenda. Each of these stages has resourcing and cost implications.

The NHS has much experience on this, through the Information Standards Board.

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

   The assumption must be yes, if Government has expectations that the data will be used to inform personal choice of services, or support accountability. However, the difficulties will arise in the way they are implemented. It is likely that a common set of principles will help ensure a consistent approach across a large number of different services and organisations.

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

   At face value, there are advantages I having a single scheme for all intermediaries, but it will be difficult to design and implement.

   An accreditation scheme needs to be clear about the problems it is trying to solve or manage. There will be a need to be clear about what is to be accredited – an organisation; its separate data outputs; the publication methods; the separate uses to which they put the data, or a specific product or service. If it applies to uses, there will be a need to be clear about the user's intent, and the potential for using data for the purpose of jigsaw reidentification is a real issue.

   The model outlined on page 26 is presented as one for continuous improvement for an organisation, but it actually works better as a way of signalling the level of sophistication associated with the method of publication for each data set. It demonstrates the need to decide what exactly is to be "accredited".

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.**

   As noted in this response, and in other reports, there is a need for transparency about transparency, especially where it relates to the different interests involved. There is a need for a high level description of the principles that must apply – especially in regard to the management of risk and opportunities for risks to emerge some time after publication (eg use of different data sets for jigsaw reidentification).

   If the high level principles are sufficiently clear, the onus can be on the organisation publishing the data (rather than the user) to ensure that the relevant considerations have been made, and there is an audit trail that might be published alongside the data.

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

This is unlikely to be straightforward, as it will need to take account of the fact that the public are likely to access much of the data in a useable format via intermediaries.

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

   It is reasonable to expect that Government will want some form of sanctions which should be clear at the outset. Clarity of principles, definitions and requirements associated with Open Data would go a long way to reduce the need for them to be invoked.

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

   We do not offer a response to this question, but assume there will be some consistency of form and function across the Boards.

Meaningful Open Data

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

   It is necessary to agree the scope of the inventory, and its intended audience/use. The form of an inventory (and its resourcing implications) will also depend on the extent to which individual organisations manage their own inventories or they are linked into a single public service inventory.

   It is possible that an incremental approach may be necessary.

   The value of an inventory will be as good as its accuracy. Arrangements for its ongoing maintenance need to be clear at the outset.

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

   Both these questions are predicated on a number of assumptions that the consultation is seeking to address, and therefore we do not offer a response to this.

   We have commented elsewhere on relevant issues, such as the need for clearer definitions of data, the difficulties associated with attributing value based on the range of diverse interests involved, and the need for a clear purpose of the inventories, etc. There will be a need also to consider cost implications.

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

   If the principle of “presumption to publish” is to be applied, then it should be easy t limit exceptions, presumably based on risk of identification and an inability to demonstrate value for money.

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

   There is already experience across the public sector on this which offers a range of learning points. For the NHS IC it is an essential part of our role that we show due regard to the administrative burden associated with collection of data.

   However, both these questions are predicated on a number of assumptions that the consultation is seeking to address, and therefore we do not offer a response to this.

   The key issue concerns the need to take account of all users of the data. We have commented elsewhere that there is not a hierarchy of users and uses.

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 important that users of the data have some indication of the quality of the data they might use, in order to have confidence in the data and any analyses using the data. There are numerous definitions of data quality. The NHS IC believes there are numerous components, as described below (this has common currency across the NHS):

   - Accuracy - how well data reflects the reality it was designed to measure.
   - Validity – how well data conforms to agreed standards and definitions.
   - Completeness – how much data is missing.
   - Timeliness - how current are the data at the time of release.
   - Reliability – stability and consistency of data processes across collection points, and over time.
   - Provenance – how the data has been changed between capture and publication.

   There will be a trade off between making the data available as quickly as possible, compared with improved quality and / or completeness which would delay publication. The intended use of the data is relevant to that trade-off. For example, pressure to publish quickly could result in the routine publication of monthly snapshots which require a lot of processing to render the data into useful formats (eg for longitudinal analysis).

   There is a limit to how much cleansing or polishing can be done in a cost-effective way once the data has been collected. And there are significant costs associated with cleaning or improving quality of data at source – the only place it can be legitimately corrected.

Government setting 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?**

   It is not easy to offer a response to this question, as it will depend on the scope, scale and speed of publication of data. However, some thought is needed on this now, to ensure that there is a consistent approach across relevant organisations, and they do not duplicate effort.

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

   The factors will be different at national, local and sector level, but there must be some principles which are applied consistently (utility, value for money, etc).

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

   We do not offer a response to this question – they are not mutually exclusive, and it depends on the strategic objective. Each case needs to be taken on its own merits – as is the case now. There will be costs associated with this, however, as it is likely to involve publishing the data in formats which are different to the formats used for collection purposes.

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?**

   We do not offer a response to this question, as it is predicated on the Government having an interest in the different uses to which the data might be put. We note, however, that there is good practice from other Government activities, involving the use of grants and awards, educational or showcase events.

2. Consultation on a Data Policy for a Public Data Corporation (PDC)

Charging for PDC information

1. How do you think Government should best balance its objectives around increasing access to data and providing more freely available data for re-use year on year within the constraints of affordability? Please provide evidence to support your answer where possible.

We have already made significant amounts of data available for public use via our own website, and via www.data.gov.uk. On the basis of our experience on the Open Data agenda so far, it is clear that:

- There are costs associated with the acquisition and processing of the data;
- There are costs associated with the current publication of the data for specific purposes for which the NHS IC has been commissioned;
- There will be additional costs associated with the publication of data for the purposes of the Open Data agenda;

The consultation document has noted that greater access to public data will require investment in infrastructure to make access possible at the scale and spend required. Our experience confirms that different factors affect the costs associated with publication of different datasets, including:

- Data currently available only to the NHS;
- Granularity of published data not adequate for the purpose of the Open Data agenda;
- Data processing requirements to render the data reusable;
- Clarity about what is published, including updating arrangements (over and above the publication of monthly updates;
- The need for additional processing, eg statistical disclosure controls especially in regard to small numbers;
- The extent to which supporting tools or documentation might be required to assist users (eg understanding of the data for analytical purposes);
- The relative ease of access to reference data such as organisation codes;
- The extent to which the reference data and granularity is consistent with the likely uses to which the data will be put.

Decisions about balancing the obligation to publish more data with affordability constraints will need to take account of a number of factors:

- Clarification as to whether the “obligation to make more data freely available year on year” means updates of data already published, or the publication of new data, or both. How will the baseline be set, and what will be the monitoring arrangements;
- Decisions about utility and value;
- There will be discrepancies between:
    - The organisations which hold the source data;
    - The organisations funding the publication;
    - The organisations responsible for making the data available;
    - The organisations re-using the data for commercial or other purposes;

- The consumers of the data or products using the data.

It is not clear that a balance can be achieved. There will be a need for trade-offs. For instance, it is likely that the Government's interest in stimulating the information market place will require some early incentives for investment, and these may challenge other interests, at least in the short term. As technologies adapt, and demand is clearer, there may be less of a need for incentives.

2. Are there particular datasets or information that you believe would create particular economic or social benefits if they were available free for use and re-use? Who would these benefit and how? Please provide evidence to support your answer where possible.

The immediate priority for the Open Data agenda is to stimulate demand and a market. Top of the list for publications therefore should be those reference data sets which would be used by others to link and cross-refer data – eg organisation codes, locations, services, etc. Beyond that, it is premature to make assumptions about interests, demand, or utility. A range of practical issues will affect the use of datasets – including publication formats.

It is therefore better to adopt some general principles about use, rather than assume any particular interests or benefits. Moreover, given the different perspectives involved, it is our view that it is not appropriate to assume a hierarchy of interests in terms of benefits which might be delivered, or to offer comments on specific datasets or the benefits they might deliver.

This is consistent with our current approach is to review all the data we hold with a view to identifying datasets for publication (applying the "presumption to publish").

3. What do you think the impacts of the three options would be for you and/or other groups outlined above? Please provide evidence to support your answer where possible.

We do not offer a response to this question.

As noted in the consultation document, it is essential that there is transparency in regard to the charging arrangements which are introduced.

4. A further variation of any of the options could be to encourage PDC and its constituent parts to make better use of the flexibility to develop commercial data products and services outside of their public task. What do you think the impacts of this might be?

Some services will be better placed to take advantage of such an opportunity, but this is an issue where there is a need for consistency across all organisations with responsibility for public data. The main impacts reflect different interests – it may help ensure that data is available in a format suitable for general public use, but it will confuse the market, and may deter investment.

5. Are there any alternative options that might balance Government's objectives which are not covered here? Please provide details and evidence to support your response where possible.

We do not offer a response to this question.

Licensing

6. To what extent do you agree that there should be greater consistency, clarity and simplicity in the licensing regime adopted by a PDC?

It is possible that the license arrangements may be a bigger barrier than cost, so the arrangements must be clear.

If the "Open Data" principle is the default position, then it should be possible to put more of the onus on public services to identify those information items where some form of license cover is required. Potential users should be able to act on the basis that they are able to reuse any data which has been made publicly available.

Ultimately, all three are equally essential for ensuring that there is awareness and understanding of the arrangements. There is also a need to identify those principles which might apply more widely across public sector organisations in regard to publication under transparency agenda.

However, it is not possible to be too prescriptive at this stage as the full range of organisations & services which might be included in the PDC is not yet clear.

7. To what extent do you think each of the options set out would address those issues (or any others)? Please provide evidence to support your comments where possible.

We do not offer a response to this question, but note that objectively, option 2 probably offers the best way of balancing principles and flexibility.

8. What do you think the advantages and disadvantages of each of the options would be? Please provide evidence to support your comments

We do not offer a response to this question.

9. Will the benefits of changing the models from those in use across Government outweigh the impacts of taking out new or replacement licences?

We do not offer a response to this question.

Regulatory oversight

10. To what extent is the current regulatory environment appropriate to deliver the vision for a PDC?

It is not possible to offer a response to this question, as there are different arrangements applying to different aspects of public data. It is only when the long term vision for the

PDC has been clearly articulated that we would be able to review the regulatory environment that is to apply to the PDC and to organisations not in the PDC.

11. Are there any additional oversight activities needed to deliver the vision for a PDC and if so what are they?

As noted above, it is likely that the establishment of the PDC will bring changes that affect a larger number of organisations, not just those in the PDC. It will need to address any issues which come to light involving organisations not in the PDC.

Moreover, the organisational form of the PDC will determine the regulatory and oversight requirements.

Therefore, at this stage it is not possible to offer a response to this question.

12. What would be an appropriate timescale for reviewing a PDC or its constituent parts public task(s)?

This may change over time. In the first instance, it is probably best to review annually for the PDC as an entity – especially if there is an assumption that it will bring in new organisations or services. It may be possible over time to move to a situation whereby the constituent parts are reviewed less frequently - it is not clear that there will be sufficient changes that warrant an annual review.

Tim Straughan
27th October 2011