Towards a pan EU data portal – data.gov.eu

Professor Nigel Shadbolt

nigel_shadbolt@gmail.com
15th December 2010
Version 4.0

1 Executive Summary

1. There is a rapidly growing move towards the release of open data. EU Member States are in the vanguard of this movement with countries such as the UK putting large amounts of its Public Sector Information on line.

2. Publishing Public Sector Information (PSI) promotes transparency and accountability, supports evidenced-based policy, creates social and economic value, achieves efficiencies and can improve the quality of data itself.

3. As this happens we need to build data portals – sites like data.gov.uk catalogue the data being released and offer a range of additional services. These portals act as unified points of access to find the published data.

4. data.gov.eu would be an EU portal that would aggregate national efforts. Data from across the EU would be catalogued and discovered using common standards. It would showcase applications that exploit the data. It would connect communities of developers and users, organisations and individuals, private and public bodies from across the EU as they use and exploit the data.

5. Challenges include a scarcity of PSI in some member states – data.gov.eu could act as a stimulus here. There are also well-understood financial, legal and organizational challenges around the publication of PSI. Political leadership and Public Data Principles can deal with these emphasizing the advantages and benefits of PSI publication.

6. Technical architectures exist that would make the implementation of data.gov.eu relatively straightforward. A modest programme of resource could deliver the portal. If procured and managed in an agile fashion as a Beta site (one under continuous refinement) it ought to be deliverable in 6 months at a cost of 500K€ - thereafter maintenance of the basic site at around 200K€ per year – depending on the speed at which national data became available.

7. Commissioning an initial project in this area need not be expensive, would provide valuable services, deepen our understanding of how to publish and exploit EU PSI, build a community of users and developers capable of creating and deriving value from the data across all EU Member States.

2   Introduction

A number of EU Member States (e.g. data.gov.uk in the UK) and elsewhere (the US with data.gov) have developed or are now in the process of creating open data catalogues and portals of Public Sector Information (PSI). This approach is also beginning to be implemented in a limited number of municipalities and regions (e.g. Catalunya dadesobertes.gencat.cat, Piemonte dati.piemonte.it). These initiatives are being undertaken by the public sector and are providing access to a broad and substantial set of government data, making data accessible in electronic 'raw' data formats (e.g. XLS, CSV) allowing for its immediate re-use. These initiatives have attracted widespread interest and are highlighting how public data can be made more accessible and available for reuse. The setting up of a data.gov.eu, an EU portal aggregating national portals (UK, FR, ES, etc.) could become an EU flagship initiative allowing governments, companies and citizens to easily find, understand, and re-use data created and maintained by the European institutions and Member States. This report is a scoping paper intended to offer guidance as to what the initiative could offer, the services it might provide, as well as the problems to be surmounted and a preliminary tentative project plan to implement a working system.

Specifically the report addresses the following:

1. It determines the potential European added value of such an initiative, identifying potential benefits to European re-users and citizens

2. It identifies the different potential types of services the EU portal could/should eventually provide.

3. It identifies the main technical, financial and organisational challenges the establishment of the portal would entail

4. It presents a tentative timeframe and proposes intermediary steps/milestones (feasibility study, pilot project, etc.)

3   data.gov.eu motivation

The past 18 months has witnessed the emergence of a number of open government data (OGD) initiatives. Two of the most developed are to be found in

the UK and the US. However, Member States, regions and cities are also embracing open data initiatives[1]. The motivations vary but a number of benefits have been advanced for these OGD projects.

The first of these is **transparency** – simply put the provision of detailed information relating to European common interests such as taxation, spending, education, transport, energy, environment, crime, health etc. enables citizens to be better informed and to be able to make comparisons within and between states.

Related to this is **accountability** – information in these sectors holds the providers of public services accountable – from spending on infrastructure to the timeliness of transport, from death rates in hospitals to employability of graduates.

Open data is also the basis for **evidence-based policy** – this provides EU institutions and Member States with the ability to base their policy decisions on empirical data – data that is open to public scrutiny and debate.

The provision of open government data across Member States can also generate **social value** – social value arises from the public good to which the data can be put. This can range from community action to remedy noise pollution to equitable access to public transport, from identifying those disadvantaged by poor IT infrastructure to providing the means for patients to share insights and experience.

As well as social value open government data can give rise to **economic value** - open public data can be aggregated and enriched to offer new services that can generate real economic returns – examples include looking at spending patterns to determine intelligent procurement, building smart transportation applications from publically aggregated time-table data across Europe. These applications are built outside of government procurement and exploit data in ways not anticipated by the collecting agency.

---

[1] Public Sector Information (PSI) Data Catalogues (by governments)
http://www.epsiplus.net/psi_data_catalogues/category_1_public_sector_informa

Open data also offers the opportunity to achieve **efficiencies** throughout public services and government, between and within Member States. Examples range from more efficient procurement to reducing the costs of servicing Freedom of Information requests, from cutting the cost of publishing data in a myriad of non-machine readable formats to reducing the cost of locating and retrieving data from across Member States. Open data can support more for less.

There is also growing evidence that open data can **improve data quality** – public data is often incomplete, out of date or of variable quality. Crowd sourcing data improvement is possible once data is openly published and a feedback means provided. Whether it is the location of bus stops or potholes in the street – the public provides a powerful resource to locate, identify and report the actual facts of the matter.

Above and beyond these general benefits there are a number of particular potential benefits that would accrue from a pan European data portal.

- Provides a single point of access for European public sector data.
- Promotes a "race to the top" as knowledge of successful OGD initiatives stimulates new initiatives in other Member States.
- Supports interoperability across processes thanks to the greater availability of data.
- Provides better comparability of EU 27 information and data both between Member States and for the purposed of wider global comparison.
- Reduces EU administrative costs.
- Eliminates or cuts down re-publication costs of official EU information.
- Provides planning and monitoring resources for those companies that operate across EU borders.
- Supports the European innovation process.
- Facilitates the harmonisation of standards and guidelines for open government data across Europe.

4   data.gov.eu services

In this section we review the types of service such a portal could provide. Two classes of service would be central to data.gov.eu – firstly the *locator services* that

use metadata to catalogue and discover relevant content and secondly *records creation* that support the creation and deposition of data assets into the portal.

One might assume that Governments and their agencies maintain the data. This is current practice at already established portals – e.g. data.gov.uk (See Appendix A). However, in the mix of practice we see proposals for the development of server "clouds" to outsource datasets. Thus cloud-hosting data at EU level might change some of the assumptions below. We are also witnessing a number of non-governmental agencies and groups aggregating national datasets, such efforts but not explicitly supported by national governments (e.g. in Ireland opendata.ie).

4.1   Locator Services

When referring to locator services we generally mean the set of human, organisational, and technological facilities that help people find information. A high level view of the locator infrastructure is presented in figure 1. This envisages a situation where catalogues are harvested or provided by the respective Member States, and are synchronised by a mediator service (although the general architecture could also support the situation where data is harvested or supplied by agencies other than EU Member States). The role of a mediator is essential if we are to provide more than a simple catalogue of catalogues. Prior to publication, the records ought to be prepared so as to support global retrieval, multi-lingual search, and facilitate recommendation processes that utilise schematic and other alignments between the catalogues. A fundamental aspect of any portal is to support the development of a community of users capable of providing relevance feedback to the retrieval models, exploit the data in a range of applications and further enrich it by interlinking to other material.
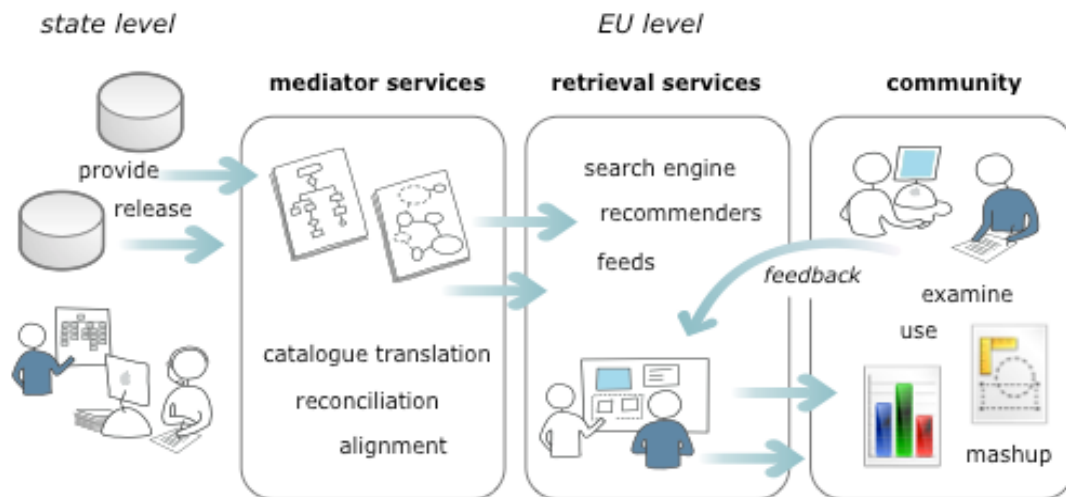
Figure 1: data.gov.eu – locator service architecture

We can see how some of these services have been provided in the largest EU Open Data project – data.gov.uk - a technical summary and topology of the physical architecture employed is provided in Appendix A. The functional components include

1. a data registry from a customised version of CKAN (http://ckan.net/) which provides an open registry of data and content packages,
2. a frontend Content Management System (CMS) using Drupal a powerful open source CMS (http://drupal.org/) (v.6),
3. a search engine provided by Apache SOLR an open source enterprise search engine (http://lucene.apache.org/solr/) ,
4. a Linked Data capability (http://data.gov.uk/linked-data) and hosting environment provided by Talis' (e.g. http://api.talis.com/stores/ordnance-survey/services/sparql) and TSO's Four Store services (e.g. http://gov.tso.co.uk/transport/sparql) .
5. Additional services for end-user support include blogs, wikis, forums, and mailing lists.

CKAN is the main mediator service of data.gov.uk and the primary route for publishers to include their data into the catalogue. The great majority of datasets remain hosted on the publisher's site, not data.gov.uk. The architecture is further split into two main subnets: a subnet used for primary connections, and a backup subnet for fail-over connections (more details are provided in Appendix A).

A fundamental technical challenge of data.gov.eu concerns the original formats of the catalogues fed into the mediator, their linguistic structure, and the original schema used to model them. Global schema might be the preferred option in such cases. A timeframe towards common implementation of core datasets would be desirable as would agreed taxonomies and ontologies for classifying specific elements of the catalogues.

In the following sections we outline the processes of record creation, schema design, and discuss support for Linked Data, and multilingualism. Services for end-user support and collaboration, including discovery, retrieval, and exploitation of data, are presented later in the report. Facilities for data enhancement and evaluation are also covered.

4.2   Records Creation

Creation and deposition of data records will generally occur alongside or during the period and process of collecting and manufacturing the data assets. The practice starts with the public sector body in charge of producing the data resource and ends with the release of the metadata record, and potentially the resource itself, to the online portal, or its mediator service (as in figure 1).

As the size of the catalogues expands over time, the cost and maintenance effort will increase. Keeping up-to-date and consistent thousands or tens of thousands of records will demand careful use and choice of technology. Versioning and record cleansing tools will be an important class of services and these should be addressed directly by any European-wide initiative.

It is unlikely that a central service would cover and support all European agencies throughout the entire lifecycle of their records. But such a service could prove valuable for agencies to use. Additionally, it would be useful for administering and controlling the flow of the records creation process.

Versioning tools, record cleansing and refinement applications (e.g Google Refine http://code.google.com/p/google-refine/), RDF convertors, automatic classifiers and recommenders for 'keyword' or 'category' related metadata, could be employed or unified at a central location. Promoting their use would foster better classification of records, faster and seamless integration, and an improvement of metadata elements (such as misspelled, redundant and

ambiguous tags). Further, providing a unified bundle for agencies to use, either as an integrated network-service or as a package for in-house deployment, would alleviate the task of a mediator to cleanse the records prior to publication.

5   Schema Designs

While access mechanisms, such as database technologies and user interfaces, might change over time, certain fundamental aspects will persist and carry long-term implications. Citizens searching for information resources will generally always need information to guide their search. What is the data about? Who distributes it? When and where was it accrued and published? How frequently is it updated? Successful deployment and continuous evolution of a pan-European portal will be interlinked with the ways in which such characteristics of the information are exposed to searching – we need stable schemas.

A catalogue schema should provide a common semantic basis for meaningful searching, encapsulate the salient features of information and be sufficient for user needs. At the same time, the schema must remain flexible so as to accommodate local extensions. Member States might wish to modify schema in ways that they think is best to serve their own user groups. Diversity, therefore, might be commonplace, especially as the portal starts to integrate increasing numbers of catalogues, as they are made available. The role of a mediator will remain prominent for semantic mappings between one metadata perspective and another.

The w3c e-government group has undertaken work to develop a cataloguing schema for government catalogues. The outcome has been the DCAT ontology[2], and which was deployed by data.gov.uk. The vocabulary is detailed in figure 2.

DCAT proposes a unified format for publishing the contents of a government catalogue and proposes the use of several open vocabularies, such as FOAF (www.foaf-project.org/), Dublin Core (dublincore.org/documents/dces/), and SKOS (www.w3.org/2004/02/skos/). There is a clear separation between metadata and provenance information related to a dataset, the catalogue record, and the catalogue itself, which is also depicted as a unique element in the ontology.

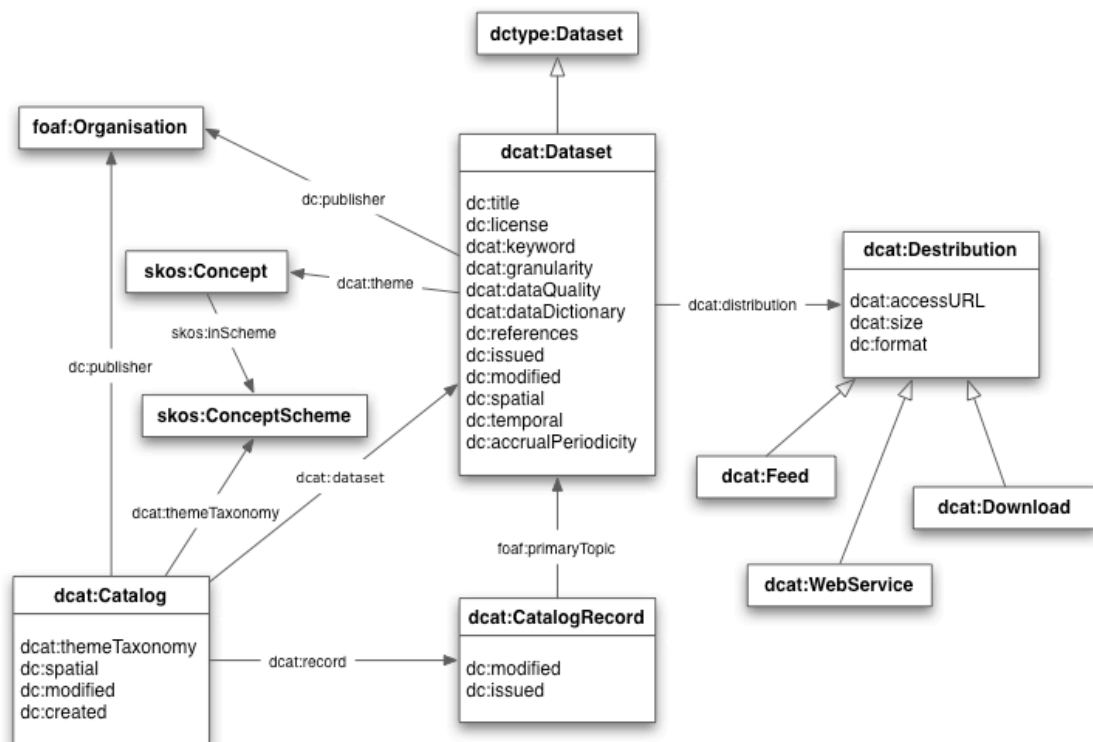[2] Data Catalog Vocabulary, DCAT, http://vocab.deri.ie/dcat-overview

Figure 2: The DCAT ontology (recreated here)

DCAT comprises a breadth of metadata, which has proved adequate for most purposes. The current metadata advice for data.gov.uk is provided in Appendix B and aligns with the DCAT ontology here.

Clearly there is a substantial amount of experience to draw on in this area. Although the UK metadata schema is itself "work in progress" it provides a useful initial design reference point. Nevertheless, there are still aspects for potential improvement and metadata schemas should be seen as subject to evolution and improvement. Some of the areas for development are summarized below.

- Although Accrual Periodicity appears to be modeled, users might also be interested in the exact methods of accrual of the data, otherwise called "collection mode".
- Fuller description of the contributors of the data: authors, publishers, third parties involved, and maintaining provenance trails when the responsibility for data collection changes.

- Information about the cost of creation and dissemination, and liabilities of use (if not covered by any open data licenses used).
- Connections to a taxonomy for representing geographical boundaries, such as for *dc:spatial* or *coverage.* This will enable seamless topological integration of the records, an important factor for cross-state search and retrieval. Related work has been ongoing as part of the NUTS[3] hierarchical classification for the economic territories of the EU. A more recent initiative at the University of Southampton has integrated the NUTS taxonomy with RDF and Linked Data formats[4].
- Need to represent meta data relating to large statistical datasets. There is existing work to publish large-scale statistical datasets as RDF and Linked Data[5]. This has arisen from efforts to integrate existing statistical metadata schema such as SDMX with the W3C Linked Data standards.

The proposed classification schemes for essential elements in the schema should ideally be adopted over time by the local agencies. For example, a uniform format for temporal information – ISO-8601, or the aforementioned NUTS taxonomy. This will, again, alleviate the task of the mediator and reduce possible ambiguities. Considerations here may be aligned with those discussed in section 3.2 on *Records Creation,* since part of the proposal is to find pathways to assist agencies in the correct delivery of records. It is important to promote the widest possible application of the schema at the point of creating the records.

6   The Road to Stardom - Supporting Linked Data

In promoting the publication of Open Government Data there have been a variety of practices to date. Across the EU Member States it is likely that there will be a wide variety of formats. In order to succeed any EU data portal should accept all structured data formats.

---

[3] Nomenclature of territorial units for statistics, NUTS, Eurostat, http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction
[4] Gianluca Correndo, Manuel Salvadores, Yang Yang, Nick Gibbins, and Nigel Shadbolt. "Geographical service: a compass for the web of data". In WWW2010 *Workshop: Linked Data on the Web* (LDOW2010), April 2010.
[5] http://data.gov.uk/resources/coins

Nevertheless best practice suggests the adoption of a methodology that supports progression to more useful formats in terms of data integration and enrichment.

The 'five star' methodology outlined by Sir Tim Berners-Lee[6] has been adopted within data.gov.uk – this outlines a pragmatic means to provide for a progression through increasingly more flexible formats.

★	if the data is available on the Web (any format)
★★	if the data is made available as structured data (e.g. the xls output from Excel)
★★★	if the data is made available using open, non-proprietorial standard formats (e.g. the comma separated values CSV format)
★★★★	if URIs are used to identify those things the data represents (so people and machines can point at your data)
★★★★★	if the data represented as in 4 is linked to other people's data (Linked Data approach)

A feature of data.gov.uk has been the adoption of linked data principles. Whilst a majority of datasets are submitted as 3 star data the UK effort has seen the construction of a "digital estate" that comprises a set of URIs describing the objects of interest. The paradigm of Linked Data is a principled approach to connecting structured data on the Web using *typed links* i.e. links that enclose meaning and distinct functionality.

The adoption of Linked Data standards for open data is growing rapidly[7]. There has been significant adoption by large organisations such as the BBC, the Library of Congress, the Guardian, and other PSI portals – the US equivalent data.gov and UK's data.gov.uk. It is envisaged that in time all these will be linked together, in ways that will allow users to start browsing in one data source and gradually navigate along links to related data sources. The development of data.gov.eu could represent a leading initiative towards this goal. The adoption of Linked Data principles to federate European catalogues with similar initiatives around the world will pave the way to the next generation locator services.

---

[6] Linked Open Data Star Scheme, http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/
[7] http://linkeddata.org/ and http://en.wikipedia.org/wiki/Linked_Data

Ownership of the URIs on the Web of Data (sometimes expressed simply by their namespace) is crucial as is provenance and persistence. This applies not just to the domain content within the datasets but also to the metadata and catalogue URIs. Should record and metadata URIs point back to the portals of the member states, or should they be centralised at the EU platform i.e. be modeled with the EU namespace? This is a decision that will drive some parts of the general architecture.

Whatever the decision it highlights a potentially important pair of roles for a pan EU portal.

1. An authoritative registry of conceptual models (e.g. the definition of a terms in domains such as transportation or spending). This would facilitate analysis, exploration and application development across data sets contributed by several data providers.

2. An authoritative provider of codes, identifiers, URIs (e.g. to see what this might look like see the URIs for administrate geography provided by the UK Ordnance Survey http://data.ordnancesurvey.co.uk/) that could be reused by national initiatives when structuring their own data. This would make it possible for information about the same entity to be retrieved across independently developed datasets.

7 Multilingualism

Linguistic and cultural diversity is a key feature of European identity; it comprises the shared heritage of an integrated Europe but presents a challenge to media and content online. The European Union is a continent of many languages. There are currently more than 23 official languages in the EU, with 60 or so indigenous, regional and minority ones.

While the learning and dissemination of European languages is encouraged, multilingualism in many digital communication technologies remains a challenging task. Work on multilingual lexicons and research on language-independent retrieval models are beginning to receive attention. Examples are the increasing availability of electronic lexicons and translators, provided as

open source and using international standards, such as the EuroWordNet database[8].

Member States should have different choices for the languages of their catalogues and, the terms and elements within their datasets. However, then we face the challenge of catalogue translation between the different European languages. Multilingual support would be a prominent asset of data.gov.eu but at the same time it represents an enormous challenge to overcome.

Several questions need to be addressed. How many languages should the portal support? Which parts of the catalogues are most essential for translation to reach a minimum functional stage? What are the most effective ways to use technology to implement multilingual support in the portal? Can we reduce complexity by coordinating a manual translation phase?

There is a case to be made in the first instance for a "hub and spoke" translation methodology. For catalogue schema and the associated taxonomies we would require that the initial creation of the record also supplied a translation mapping of the terms into English. This would enable core search methods to map into the native datasets in a relatively straightforward manner. More sophisticated translation support could be developed subsequently.

8   Search and Retrieval Services

During the last 30 or so years, work on information retrieval has grown beyond its primary goals of indexing and searching for documents in library collections. Increasingly we see the importance of intelligent ranking algorithms to improve the quality of answer sets, studying the behaviors of users and understanding their needs, delivering improved graphical interfaces and collaborative frameworks.

Several technical approaches have been pursued; relevance feedback to embody user satisfaction about the retrieval process, exploiting document semantics, query log analysis to correlate user requests, query expansion via local and global graph analyses, use of term association and metric clusters, development of similarity and statistical thesauri. If we take account of this rich range of

---

[8] http://www.illc.uva.nl/EuroWordNet/

information retrieval approaches we can improve user satisfaction by fulfilling the most naïve and at the same time the most complicated information requests.

The development of data.gov.eu will bring expert users and regular citizens together and require a robust and collaborative retrieval model. The rich semantic structure of government catalogues, coupled with flexible retrieval models and rich user interfaces, has the potential to deliver breakthrough services for the PSI needs of European citizens.

We outline below some of the fundamental search and retrieval functionality needed to help assist users in dealing with the volume and variety of records at such a portal. In the next section we describe additional recommendation services.

- Faceted search and ways to progressively refine search requests. Exploit and combine the rich metadata of the catalogues to provide intelligent, yet simple search utilities i.e. be able to get precise results on a guided search for hospital listings per capital or country in Europe, retrieve records about accommodation facilities and restaurants refined by country or town, etc.
- Embody user feedback directly in the retrieval/search process. There are several ways to accomplish this – some involve algorithmic approaches, such as query expansion via relevance feedback, others rely on interface features, such as iterative and interactive search. Probabilistic retrieval models also incorporate user feedback in the ranking process.
- Access to previous search requests. Browsing history and key concepts visited throughout the user's stay. Let users save their history in their profiles. Let them view their history (primarily consisting of requests and observed records) as a tag-cloud and get recommendations from it.
- Feeds for previously viewed records and downloaded datasets.
- Provide a list of datasets that cannot be released – this addresses a Freedom of Information (FOI) issue encountered in the UK and likely to exist throughout the EU. How can you ask for what you don't know is there. Asset inventories should be available with their status.
- Better tag-cloud navigation. Recent analysis of tag-clouds coming from a number of available portals – data.gov, data.gov.uk, data.australia.gov.au – reveal large amounts of agency-defined index terms, covering a wide array of topics. On average records come with about 8 such keywords, while some

records contain more than 100 pre-compiled keyword terms. These are useful and can be exploited for improving navigation of the tag-clouds.

- A multilingual retrieval service – be able to search in multiple languages.

9 Recommendation Services

A range of recommendation services[9] would be possible in a portal of this size and ambition. For example,

- *Similar users also downloaded*… Amazon-style recommendations. Some users will have a longer and more interactive experience with the portal than others. The system can provide assistance to naïve users or newcomers by correlating their profiles with expert users using features of the resources they search for, records they view, and the datasets they download. This implicit collaboration of expert with inexperienced users will certainly help individuals to mine resources that are not readily available or ranked high on their search results. Recommendations of this type may be delivered in the form of "*see what similar users are searching*", or while the user is viewing a specific record.
- Recommendations of similar datasets for locations nearby the user's specific geography. This can be employed on a per-record recommendation basis i.e. offer the recommendation while the user is viewing a specific record.
- Popular and featured datasets presented on the main page.
- Supporting network formation of "people like me" is another recommendation method that is widely and successfully used.

10 Data exploitation

The portal should feature services that enable exploitation of the data. For full engagement there are a number of distinct audiences.

---

[9] Dietmar Jannach et al, (2010) Recommender Systems An Introduction CUP, ISBN: 9780521493369

- Organisations that have the "in house" expertise to extract value from the data. For example, traditional data integrators who would find most value in the data being easy to find and available to download in standard formats and where the semantics of the data are explicit.
- Non-specialists who have an interest in the data but are not specialist programmers. Here a collaborative, simple module for putting together mashups of datasets is called for – one that allows for discussion around the data and its properties. Simple blogs and links to the datasets can suffice – see for example http://www.datamasher.org/node/106
- Developers who are moving to Linked Data and W3C recommendations[10] – here the data would need to be available as well-formed RDF with well designed URIs – either as files or available from SPARQL endpoints.
- Developers who appreciate the data in accessible forms that do not necessarily require Linked Data expertise – see for example the Linked Data API of data.gov.uk that delivers JSON and other forms[11].

Links to scripts for converting and manipulating the data should also be made available where possible. Further, the catalogues should be integrated with local data sites, even if they offer redundant means of presenting data. Available mashups and other visualisations of the data could also be linked off the main site (e.g. data.gov.uk/apps and www.data.gov/developers/showcase).

11 Data Enhancement and Improvement Services

11.1 Evaluation Services

Member States and Public sector bodies will need to undertake regular evaluations in order to determine the future strategies that should be adopted by providers of PSI These may be based around the quality, value, and quantity of records provided. The "five star" referred to earlier in this report is a proposal that offers pragmatic guidance in this progression. Further, a module with visual aids to view the evaluations will also be useful i.e. bar charts, line charts, league tables.

---

[10] http://www.w3.org/standards/semanticweb/data
[11] http://www.epimorphics.com/public/presentations/ogdc-slides/ogdc.html#(1)
http://code.google.com/p/linked-data-api/

11.2 Community Development and User Involvement

The recruitment and development of an engaged developer community was one of the major successes of data.gov.uk – the community of developers acted as "critical friends" as the system was being developed and provided much of the input to refine and enhance the look, feel, and services offered by the site.

In order to support the communities that will naturally form around the portal, obvious functionality includes

- User profiles with group membership, user networks, shared blogs, wikis, email lists and fora. In supporting these functionalities it is important to recognize the importance of active administrators and moderators.
- User ranks provide kudos for those who are the main and most valued contributors.
- A collaborative tagging module that allows users to apply their own tags to the datasets – *Del.icio.us* style
- "Ask the experts" module – discussion forums are essential elements for disseminating best practice.

It is important to remember that developers and users of PSI care about their local context first and foremost. Their primary interest is likely to be local or hyperlocal[12], regional, national, and super national in that order. Any community development needs to bear this consideration in mind. It should also pay close attention to how developers themselves are using the data.

12 Challenges

There are a number of significant challenges that present themselves when considering a pan EU portal of the sort described here. Technical challenges have been addressed already in this report although it is useful to outline additional specific items. A second set could be considered essentially financial or resource specific. Finally there exist a set of challenges around the social, organizational, legal and political aspects of a pan EU portal.

---

[12] http://openlylocal.com/hyperlocal_sites

12.1 Open Data

Perhaps the single most significant challenge is the scarcity of national EU Open Government Data initiatives. The largest national efforts are still restricted to particular English speaking democracies; UK, US, Australia and New Zealand. There are however a significant number of regional EU open data initiatives. Moreover, there are a number of EU national initiatives planned. But it is likely that initially the EU portal might have to serve as the primary data catalogue whilst national portals are commissioned. In some ways this is a strong argument in favour of the EU portal – since there are many national datasets that are ready and available for publication – but as yet no national depository exists.

12.2 Technical

The largest technical challenge is around the characterization of an appropriate portal architecture and the set of associated services. This report has reviewed the largest extant such portal – data.gov.uk - it provides an existence proof that the basic functionality can be realized. It would seem a good candidate for implementation – at least in any pilot system that was commissioned.

The particular technical problems relating to the pan EU portal are around the multilingual facilities – and there would need to be careful consideration around the initial functionality offered here. This report has argue for a minimal "hub and spoke" solution in the first instance.

It is hard enough to establish consistent data reporting standards and formats within existing national borders and administrations. These would be multiplied across the EU and there would need to be considerable political will and leadership to support this technical harmonization.

It is important to build evaluation and assessment into the architecture from the outset. This allows for the ability to measure data reuse and establishes some notion on the return on investment or savings that might result.

If linked data is to be supported or else if datasets are to be hosted by the portal there is the questions of providing sufficient computational resource. Infrastructure needs to be paid for and maintained. Large data publishers are outsourcing some of the data hosting to cloud providers. However, if SPARQL

endpoints are to be supported – allowing datasets to be queried directly over these can impose substantial computational overheads.

12.3 Financial

One financial challenge that has to be factored in by any data publisher concerns data sets that are today distributed for a fee and represent a source of income to some administrations. Some of these provide local economic benefit but the greater benefit of releasing the data to a larger audience and set of application developers has to be considered. Some data sets relating to company registration or else geography are particularly important in fusing together other data – if this key connective data is restricted by fees and licences then the benefits of open data never accrue.

The service will have a cost to set up and run. There is also a cost in terms of the capability development within publishers. People have to be given the skills to generate open data.

12.4 Legal

A substantial challenge will exist around variations and differences in legislation across Member States. For example, basic data access and storage law varies across the EU. This will make data publication a challenge for some Member States.

Another substantial challenge is striking the right balance between transparency and open data and ensuring that privacy is maintained. This might involve careful redaction of items in datasets that could identify individuals – for example in the publication of crime statistics, or spending information[13].

A fundamental precondition to the effective implementation of OGD is open licensing. Experience in many countries is that developers are discouraged from using data if there are restrictions around it. These can vary from fees that are charged, through to the assertion of "derived data" rights over the data. OGD flourishes where there are no impediments on the reuse of data. The UK Open

---

[13] http://data.gov.uk/blog/transparency-and-privacy-review-announced

Government License is one of the most significant achievements in the UK effort and serves as an interesting template[14]. Licenses need to be open for the data sets and also for the associated meta data.

An open license policy would seem to be a prerequisite for the pan EU portal and essential if we are to maximize reuse of data assets.

12.5 Social and Organisational

Although implicit when discussing the services needed for developers and citizens it must be explicitly recognized that there needs to be active involvement of both if we are to build viable communities of users of the data. This will require active management.

At the pan EU level it is likely that a range of suppliers of software systems and services would find the data of significant interest. The inclusion of value added material and services from private enterprise will be an important success criteria. Their involvement is important.

In building communities of engaged stakeholders an effective communications, dissemination and development programme is needed. Whether via conferences or online competitions, schools or Universities, through public or private sector bodies a feature of OGD initiatives to date has been vibrant community involvement (e.g. the Government Bar Camp[15] movement). The pan EU effort needs to replicate this.

The engagement of the media has been a particular feature in OGD. Media are increasingly looking at the concept of "stories from data" – the Guardian in the UK and New York Times in US have been two prominent proponents of this approach[16]. The increasing use of public data by information aggregators and

---

[14] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[15] http://en.wikipedia.org/wiki/BarCamp,
http://opengovernmentdata.org/camp2010/,
http://opengovernmentdata.org/camp2010/after/

[16] http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide, http://data.nytimes.com/

search engine companies also offers an opportunity for engagement and dissemination.

The organizational challenges are significant. Not least because many in participating governments will view OGD as something else they are being asked to do. The benefits have to be clearly explained and the data publishers have to be acknowledged in this process. Even with clear commitment there will be variable practice in publication capabilities. We have learnt in the UK that keeping the publication process close to those who generate the data is important. The more hubs and aggregation points that are put in the way the more difficult the organizational process.

In the case of relatively mature national OGD sites it might be simply a matter of harvesting or synchronizing with these portals. An alternative model is to ask the data publishers to submit their data to their national and the EU portal. However, one would need to be sure that no additional burden was being placed on the publishers in terms of meta data. This is another reason to find a general form of the DCAT scheme that can be adopted across the EU.

12.6 Political

There is a strong argument for trying to establish a set of Public Data Principles (See Appendix C). Many of these relate to the policy assumptions around the publication of public non-personal data. In particular embracing OGD requires the move to a position that non-personal public data will be published unless there is a clear reason not to.

Finally, and perhaps most importantly the most successful OGD initiatives have received top-level political support and leadership – in the UK successive Prime Ministers and in the US at a Presidential level. The EU is demonstrating political leadership and commitment in the statement's of EC Vice-President Neelie Kroes[17]

---

[17] http://www.epsiplatform.eu/news/news/neelie_kroes_on_eu_open_data_portal

Expectation management is important and at the same time that a pan EU portal is being built the Commission's own data should be made available. Not "do as I say" but "do as I do".

There is great value in publishing all the data sets one can find – on the principle that others will find applications for data that governments had never thought of. This unanticipated reuse of content was one of the major lessons of Web 1.0. However, public interest and enthusiasm is maintained when "high value" datasets are released – data that the public cares about. In the UK this has been around how the Government spends taxpayers money, crime and health data, data relating to schools and transport etc. One should aim in pan EU work to surface such data. Indeed one of the reasons for having such a portal is that it produces an environment in which citizens of one Member State can ask why if others can have certain public data released why can't they.

13 Timeframe and Deliverables

Despite the complexities of a pan EU portal given the rapid developments at national and regional level there is a strong argument for launching a pilot project quickly.

However, one of the immediate challenges that follows on from the current lack of EU national portals is establishing whether to go for a broad based catalogue – where the portal acts as an official depository whilst national efforts get underway – or else to select a number of official or otherwise recognized national sites to act as feeds.

Whichever approach is taken a basic set of deliverables outlined below would allow for the establishment of a data.gov.eu pan EU portal. This should comprise the basic elements of the services described in this report and so would need the following as intermediate tasks, subtasks and milestones (see also the GANTT chart in Appendix D)

1. Workpackage 1 data.gov.eu data and schema design
    1. Guidelines for pan EU URI design scheme – similar in nature to the advice contained here http://data.gov.uk/resources/uris but allowing for particular design considerations from an EU perspective – 2 person months/40 days

2. Initial Data Inventory for inclusion in the portal – this will provided an overview of extant data and its respective attributes – formats, licences and other basic metadata – 2 person months/40 days

3. DCAT style schema for record creation – initial basic schema for EU data set records – 1 person month/20 days

2. Workpackage 2 data.gov.eu Open Licence

   1. Open Licence for portal content – need to determine if this was required above and beyond the licences attaching to the national data sets but certainly it would be for the pan EU portal specific data elements – 3 person months/60 days

3. Workpackage 3 data.gov.eu technical infrastructure

   1. A CKAN installation customized for EU datasets – would expect this to be in line with that maintained for example by the UK – 1 person month/20 days

   2. Basic search, retrieval and recommendation capabilities – sufficient to enable datasets to be identified, retrieved and similar items identified – 3 person months/40 days

   3. Basic minimal level of multilingual support – initially would suggest all submitted data sets additionally have a mapping of their vocabulary and data schema into English – this would provide for a hub-based mapping model that would have significant initial utility – 3 person months/60 days

   4. CMS framework – selection and design from a number of alternatives – 1 person month/20 days

   5. Basic Web 2.0 tools to support communication and collaboration – blog, wiki, posting of applications, suggestions for new data and applications, discussion forum, email lists – 2 person months/40 days

4. Workpackage 4 data.gov.eu System Integration

   1. System Integration Phase I – to ensure all elements of the system and workflow processes integrated - 2 person months/40 days

   2. Milestone Developer Release of data.gov.eu – after 40% of elapsed pilot project

   3. System Integration and Refinement Phase II – to iterate and refine elements of the portal - 3 person months/60 days

   4. Milestone Public Release data.gov.eu – 6 months after initiation of project

5. Workpackage 5 data.gov.eu Linked Data Infrastructure
    1. SPARQL endpoints – to provide Linked Data access via SPARQL endpoints running on triple store technology - 1 person month/20 days
    2. Linked Data Support Tools – modify and customize variety of linked data publication tools for the portal – 1.5 person months/30 days
    3. Core Reference Linked Data Sets - convert a number of core data sets into linked data – 2 person months/40 days
    4. Exemplar Applications Linked Data Functionality – build exemplar linked data applications using core reference data – 1.5 person months/30 days
6. Workpackage 6 Project Management and Dissemination
    1. Project Management 1 person month/20 days
    2. Dissemination and Communication 1 person month/20 days

Total pilot project resource 30 person months – or 5 full-time equivalents over 6 months – equates to around 400 K€ of staff costs (full economic cost) and 100 K€ for other costs including system procurement until launch.

Ongoing development and maintenance costs are difficult to estimate. In part because the effort required will be a function of the success of the initial deployment. If the site serves as a stimulus for Member states to publish their data then the amount of data to be dealt with could increase very dramatically – and this would have an effect on the maintenance effort and could also stimulate requests for new services. A modest ongoing cost of operation and support would be around 200K€ per year – over fives years equating to 1M€ and a total project cost of around 1.5M€.

The costs are only possible if the system is procured with the following characteristics : (i) based as open source, (ii) presented as a perpetual Beta (system under development) and (iii) adopt an open licence

This should deliver a system that could be launched within 6 months – although developers and the wider data community would be involved from the outset – so that system meets their needs. The system would have basic functionality and serve to highlight the particular challenges and opportunities afforded by data.gov.eu

Based on national experience building such sites can happen if procured and managed using agile methods. This is best served by a small team of dedicated developers with a few officials having the political backing and authority to surface the initial data sets.

14 Conclusions

There is a significant momentum behind OGD. There are clear economic, social and political benefits. Within the EU Member States there is a range of experience and some developed national expertise. The number of official Member State portals is still small but more can be expected in arise in the course of the next year.

If the EU does nothing there are still likely to be efforts to catalogue the OGD efforts. However, as this report and an earlier Workshop[18] points out a pan EU portal could act as a stimulus for national efforts. This report considers that commissioning an initial project in this area need not be expensive, would provide valuable services, deepen our understanding of how to publish and exploit EU OGD, build a community of users and developers capable of creating and deriving value from the data across all EU Member States.

15 Acknowledgements

This report reflects the personal view of the author. He gratefully acknowledges input from James Forrester of the data.gov.uk team and Christos Koumenides from the University of Southampton. The report also benefited from a workshop hosted by the Commission at which the author was present.

---

[18] http://cordis.europa.eu/fp7/ict/content-knowledge/docs/report-ws-pan-eu-dat-porta_en.pdf

Appendix A An Open Government Data Service Deployment: the data.gov.uk example at December 2010

Architecture

There are four primary functional components of data.gov.uk's main Web service – (i) the data registry, (ii) front end, (iii) search, and (iv) the Linked Data hosting. Other than Linked Data and very occasional use e.g. for the HM Treasury Comprehensive Spending (COINS) datasets, hosting of datasets is provided by the publisher, not data.gov.uk
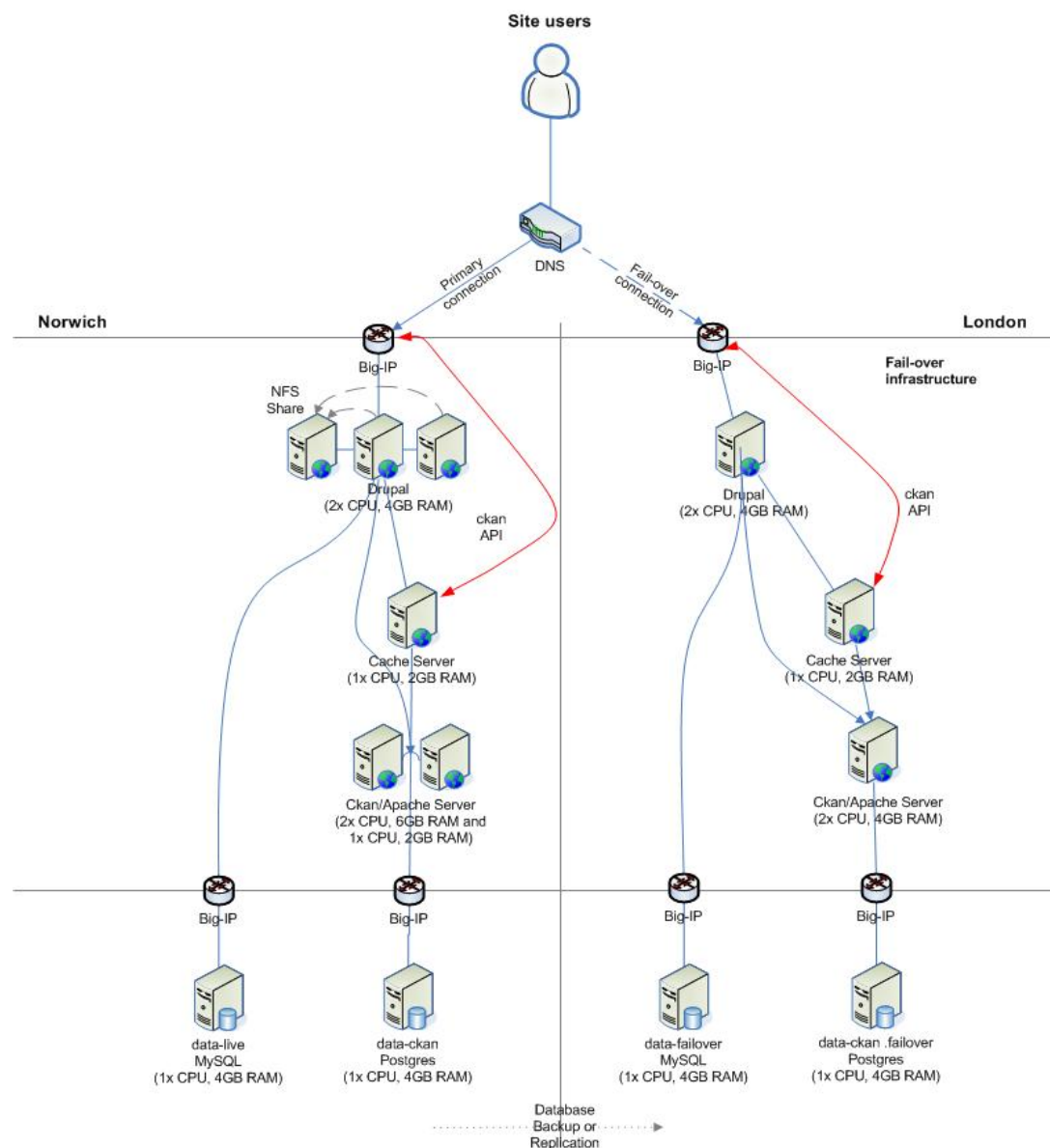
data.gov.uk as-is physical architecture (excluding Linked Data services)

The data registry is provided from a slightly-customised version of CKAN which provides the open registry of data and content packages. The customized version both leads and lags from the public version running ckan.net, and is subject to on-going development (especially with regards to implementing the EC's INSPIRE directive). Code changes are released into the main CKAN tree when they are stable.

The service includes a (read-only) public API at *catalogue.data.gov.uk* which returns JSON, JSON-P or using the CKAN client library. There is an RDF view of this catalogue. The RDF service is an "official" service, but it is not native, instead it originates from an RDF-converted daily dump from the UK Government CKAN. It is available from the API, e.g. http://catalogue.data.gov.uk/doc/dataset/coins). There is an argument for building a native RDF-enabled service based around triple store technology rather than the current Postgres RDBMS.

The frontend CMS is provided using Drupal (v. 6). Meta-data about the datasets is mastered in and provided from CKAN directly except for a few items (e.g. user ratings, links between datasets and related apps, etc.), which will be moved over to CKAN as and when appropriate.

There is RDFa in selected areas of the frontend (e.g. dataset description pages), but a comprehensive version with the human- and machine-readable versions produced from the same back-end would require moving to a different CMS (assumed to be Drupal 7, when this becomes stable).

Search is provided by Apache SOLR (open source enterprise search platform), mastered by CKAN and also written to by Drupal. This is surfaced through Drupal for the human- readable version; a machine-viewable form is not yet public.

Linked Data hosting is provided using Talis's and TSO's Four Store services with Puelia (the Linked Data API) supplying the joint human-/machine-readable interface on top – see e.g.:
http://education.data.gov.uk/doc/school?parliamentaryConstituency.label=Islington%20North

Core services

Services specific to the data catalogue.

Current data catalogue services:
Reading

- Human- facing (via Drupal and CKAN)
- Machine- facing in the API (via CKAN – some elements not yet available)

Writing

- Human- facing (via Drupal and CKAN)
- Import scripts from the Office of National Statistics (ONS) and Data4NR; occasional bulk-imports

Searching

- Human-facing structured search (from SOLR via Drupal)
- Machine- facing in the API (from CKAN – not structured)

Linked Data

- Access to SPARQL endpoints for key datasets
- API for Linked Data

Communication

- Blogs and best practice advice and guidance (including multimedia and video
- ideas and application catalogues – to harvest ideas for new data sets and showcase applications using the data
- wikis to support discussion of range of topics around methods, tools, techniques and datasets
- fora and mailing list

Potential future services:

- Devolving the data catalogue writing via API to data publishers.
- Data access (not just meta-data) via catalogue API – i.e. proxying of payload on request.
- Comprehensive RDF forms of the data catalogue – requires back-end re-write of CKAN.
- Data catalogue search via the API to use SOLR rather than plain-text search.
- Transfer and presentation of the community-focussed meta-data from Drupal to CKAN.

Workflow

Datasets are published into data.gov.uk's catalogue through three main routes:

- The primary route for publishers to include their data into the catalogue involves using a Web form for authenticated users through the Drupal data.gov.uk system into CKAN via an API. There is currently no provision for pre-moderation of data releases made in this way, and The National Archives (TNA) go through the published data and follow-up with data publishers to improve the meta-data. TNA also undertake the addition to the catalogue of data from some publishers.

- An additional route are import scripts that take feeds from the Office for National Statistic's Publications Hub and the Department for Community and Local Government's Data4NR[19] service, and occasional manually-driven bulk-imports of data from large scale data publishers.

- Linked Data work done by the team at The National Archive (TNA) has a different route that involves peer review of the modelling and representations of data before publication. Not all such data is included in the data.gov.uk index, but this will change soon.

Datasets provided under the EC's INSPIRE Directive will (by law) be mass-submitted via a different set of routes (mainly CSW), which is under development and will be deployed over the next few weeks and months.

---

[19] http://www.data4nr.net/introduction/

Appendix B Description of meta-data elements: the data.gov.uk example

The instructions below were developed by the data.gov.uk team to help public bodies supply the information required to include their published datasets in the data.gov.uk Internet service.

Given the large amounts of data that the UK public sector will provide through data.gov.uk, the datasets must be well-catalogued. It is essential to provide each dataset with a carefully constructed record that facilitates user activities such as discovering, searching, linking, and comparing datasets.

The elements of meta-data have been designed to provide flexibility for cataloguing any public datasets; guidance for each element is given below.

Dependent on what kind of data-set is supplied, some elements are "**Mandatory**" and must be completed. This appendix provides an at-a-glance view of which elements are required under each "profile".

There is the option to add others if helpful for members of the public.

It is important that publishers understand that all data submitted in this form will be published on the Internet and must be UNCLASSIFIED and in some cases a process of redaction will be required.

The project team for the data.gov.uk work can be reached at publicdata@nationalarchives.gsi.gov.uk.

Identifier

This is a public unique identifier for the dataset. It should be roughly readable, with dashes separating words. If the data relates to a period of time, include that in the name. Indicate broad geographical coverage to distinguish from those datasets for another country.

Format	Two or more lowercase alphanumeric or dashes (-).
Example	uk-road-traffic-statistics-2008 *or* west-midlands-employment-statistics-2010
Obligation	**Mandatory** for all datasets on data.gov.uk
Change	In the previous meta-data guidance this was called "Name".

Title

This is the title of the data set so that the public can easily tell what the dataset covers. It is not a description. Do not give a trailing full stop.

Example	Road traffic statistics for major UK roads, 2008
Obligation	**Mandatory** for all datasets on data.gov.uk

Abstract

This element is the main description of the dataset, and often displayed with the package title. In particular, it should start with a short sentence that describes the data set succinctly, because the first few words alone may be used in some views of the data sets.

Obligation	**Mandatory** for all datasets on data.gov.uk
Change	In the previous meta-data guidance this was called "Notes".

Date released

This is the date of the official release of the <u>initial version</u> of the dataset (this is probably not the date that it is added to the data.gov.uk index). Be careful not to confuse a new 'version' of some data with a new dataset covering another time period or geographic area.

Date updated

This should be the date of release of the <u>most recent version</u> of the dataset (not necessarily the date when it was updated on data.gov.uk). As

with **Date released**, this is for updates to a particular dataset, such as corrections or refinements, not for that of a new time period.

Date to be published

The date when the dataset will be updated in the future, if appropriate.

Change   New for this version of the meta-data guidance.

Update frequency

This should be how frequently the datasets is updated with new versions. For one-off data, use "never". For those once updated but now discontinued, use "discontinued".

Choices   Never | Discontinued | Annual | Monthly | Weekly | *Other as appropriate*

Precision

This should indicate to users the level of precision in the data, to avoid over-interpretation.

Example   per cent to two decimal places *or* as supplied by respondents

Geographic granularity

This should give the lowest level of geographic detail given in the dataset if it is aggregated. If the data is not aggregated, and so the dataset goes down to the level of the entities being reported on (such as school, hospital, or police station), use "Point". If none of the choices is appropriate, please specify in the "other" element. Where the granularity varies, please select the lowest and note this in the "other" element.

Choices   National | Regional | Local Authority | Ward | Point | *Other as appropriate*
Example   For a list of all schools in a Local Education Authority's area, "Point".

Geographic coverage

This should show the geographic coverage of this dataset. Where a dataset covers multiple columns, the system will automatically group

these (e.g. "England", "Scotland" and "Wales" all being "Yes" would be shown as "Great Britain").

Choices   Yes | No *for each.*

Temporal granularity

This should give the lowest level of temporal detail given in the dataset if it is aggregated, expressed as an interval of time. If the data is not aggregated over time, and so the dataset goes down to the instants that reported events occurred (such as the timings of high and low tides), use "Point". If none of the choices is appropriate, please specify in the "other" element. Where the granularity varies, please select the lowest and note this in the "other" element.

Choices   Year | Quarter | Month | Week | Day | Hour | Point | *Other as appropriate*

Example   For the number of enquiries dealt with each day, "Day".

Temporal coverage

The temporal coverage of this dataset. If available, please indicate the time as well as the date. Where data covers only a single point in time, give the instant rather than a range.

Example   21/03/2007–03/10/2009 *or* 07:45 31/03/2006

URL

This is the Internet link to a web page discussing the dataset.

Example   http://www.somedept.gov.uk/growth-figures.html

Taxonomy URL

This is an Internet link to a Web page describing for re-users the taxonomies used in the dataset, if any, to ensure they understand any terms used.

Example   http://www.somedept.gov.uk/growth-figures-technical-details.html

Ministerial Department

> This is the sponsoring ministerial Department under which the dataset is collected and published (but not necessarily directly undertaking this – use the **Organisation** element where this applies). The data.gov.uk system will automatically capture this.

> *Obligation*   **Mandatory** for all datasets

Organisation

> This is the body responsible for the data collection, if different from the **Ministerial Department**. Please use the full title of the body without any abbreviations, so that all items from it appear together. The data.gov.uk system automatically captures this where appropriate.

> *Example*   Environment Agency

Publisher

> An "over-ride" for the system, setting the public body that should be credited with the publication of this data, instead of the **Organisation** or **Ministerial Department**. This could be used where the public branding of the work of an agency is as its parent department.

> *Change*   New for this version of the meta-data guidance.

Contact

> This is the permanent contact point for the public to enquire about this particular dataset. This should be the name of the section of the agency or Department responsible, and should **not** be a named person. Particular care should be taken in choosing this element.

> *Example*   Statistics team | Public consultation unit / FOI contact point
> *Change*   In the previous meta-data guidance this was called "Author".

Contact e-mail

> This should be a generic official e-mail address for members of the public to contact, to match the **Author** element. A new e-mail address may need to be created for this function.

> *Change*   In the previous meta-data guidance this was called "Author e-mail".

Mandate

> An Internet link to the enabling legislation that serves as the mandate for the collection or creation of this data, if appropriate. This should be taken from The National Archives' Legislation website, and where possible be a link directly to the relevant section of the Act.
>
> *Example*   http://www.legislation.gov.uk/id/ukpga/Eliz2/6-7/51/section/2
>
> *Change*   New for this version of the meta-data guidance.

Licence

> This should match the licence under which the dataset is released. This should generally be the Open Government Licence for public sector information. If you wish to release the data under a different licence, please contact the *Making Public Data Public* team.
>
> *Obligation*   **Mandatory** for all datasets on data.gov.uk

Tags

> Tags can be thought of as the way that the packages are categorised, so are of primary importance. One or more tags should be added to give the government department and geographic location the data covers, as well as general descriptive words. The Integrated Public Sector Vocabulary may be helpful in forming these. As tags cannot contain spaces, use dashes instead.
>
> *Format*   Two or more lowercase alphanumeric or dashes (-); tags separated by spaces.
>
> *Example*   For NHS statistics on burns to the arms: "nhs arm burns medical-statistics"

Resources

These can be repeated as required. For example if the data is being supplied in multiple formats, or split into different areas or time periods, each file is a

different "resource" which should be described differently. They will all appear on the dataset page on data.gov.uk together.

Resource description

>   This will identify each specific resource if you have multiple ones for this dataset.

> | | |
> |---|---|
> | *Example* | 2009/10 data \| Zipped data (20 MB) \| CSV-format \| API service |
> | *Change* | New for this version of the meta-data guidance. |

Resource format

>   This should give the file format in which the data is supplied. You may supply the data in a form not listed here, constrained by the Public Sector Transparency Board's principles that require that all data is available in an "*open and standardised format*" that can be read by a machine. Data can also be released in formats that are not machine-processable (*e.g.* PDF) alongside this.

> | | |
> |---|---|
> | *Choices* | CSV \| RDF \| XML \| XBRL \| SDMX \| HTML+RDFa \| *Other* |
> | *Obligation* | **Mandatory** for each resource |
> | *Change* | In the previous meta-data guidance this was called "File format". |

Resource URL

>   This is the Internet link directly to the data – by selecting this link in a web browser, the user will immediately download the full data set. Note that datasets are not hosted by the project, but by the responsible department

> | | |
> |---|---|
> | *Example* | http://www.somedept.gov.uk/ growth-figures-2009.csv |
> | *Obligation* | **Mandatory** for each resource |
> | *Change* | In the previous meta-data guidance this was called "Download URL". |

Obligation	**Mandatory** for there to be at least one **Resource** for all items on data.gov.uk

Appendix C UK Public Data Principles

http://data.gov.uk/wiki/Public_Data_Principle

"Public Data" is the objective, factual, non-personal data on which public services run and are assessed, and on which policy decisions are based, or which is collected or generated in the course of public service delivery.

- Public data policy and practice will be clearly driven by the public and businesses who want and use the data, including what data is released when and in what form
- Public data will be published in reusable, machine-readable form
- Public data will be released under the same open licence which enables free reuse, including commercial reuse
- Public data will be available and easy to find through a single easy to use online access point (data.gov.uk)
- Public data will be published using open standards, and following relevant recommendations of the World Wide Web Consortium
- Public data underlying the Government's own websites will be published in reusable form for others to use
- Public data will be timely and fine grained
- Release data quickly, and then re-publish it in linked data form
- Public data will be freely available to use in any lawful way
- Public bodies should actively encourage the re-use of their public data
- Public bodies should maintain and publish inventories of their data holdings

Appendix D A Possible Project Plan for Pilot data.gov.eu

#	Task Name	Start Date	End Date	Duration	Pre
1	data.gov.eu data and schema design	04/01/11	06/23/11	60	
2	data.gov.eu URI design	04/01/11	05/26/11	40	
3	Identify Initial Datasets	04/01/11	05/26/11	40	
4	Design DCAT schema for Datasets	05/27/11	06/23/11	20	3
5	data.gov.eu Open Licence	04/01/11	06/23/11	60	
6	Determine Open Licence	04/01/11	06/23/11	60	
7	data.gov.eu technical infrastructure	04/01/11	09/16/11	121	
8	CKAN Installation	04/01/11	04/28/11	20	
9	Search, retrieval and Recommender tools	04/29/11	06/23/11	40	8
10	Multilingual Support	06/24/11	09/15/11	60	9
11	CMS Framework	04/01/11	04/28/11	20	
12	Web 2.0 Tools	04/29/11	06/23/11	40	11
13	data.gov.eu system integration	04/30/11	09/16/11	101	
14	System Integration Phase I	04/30/11	06/23/11	40	
15	Milestone Developer Release data.gov.eu	06/24/11	06/24/11	0	14
16	System Integration and Refinement Phase II	06/24/11	09/15/11	60	
17	Milestone Public Release data.gov.eu	09/16/11	09/16/11	0	16
18	data.gov.eu Linked Data infrastructure	04/01/11	09/15/11	120	
19	SPARQL Endpoints	04/01/11	04/21/11	15	
20	Linked Data Support Tools	04/22/11	06/02/11	30	19
21	Core Reference Linked Data Sets	06/03/11	08/04/11	45	20
22	Exemplar Applications	08/05/11	09/15/11	30	21
23	data.gov.eu Project Management and Comms	04/01/11	09/15/11	120	
24	Project Management	04/01/11	09/15/11	120	
25	Dissemination and Communications	04/01/11	09/15/11	120	