Title	Strategic approaches to Data Sharing
Revision	4
Status	Final
Date	8th April 2015
Author	Paul Davidson, LeGSB Director of Standards, and CIO Sedgemoor District Council
Purpose	To propose approaches to sharing data between central and local public sector organisations, which are repeatable, scalable, and reuse common components and standards.
Type	LeGSB White Paper

Revision	Date	Author	Notes
1	26/07/2013	Paul Davidson, for LeGSB	
2	01/07/2014	Paul Davidson	Rename 'Authorization' to 'Rights'
3	05/07/2014	Paul Davidson	Rename 'Authentication' to 'Trust'
4	08/04/2015	Paul Davidson	Remove Information Principles for the UK Public Sector 2011. Add 'Reference' as an information context.

This document

1. This LeGSB[1] White Paper proposes approaches to sharing data between central and local public sector organisations, which are repeatable, scalable and reuse common components and standards.

Background

2. LeGSB has developed a 7 theme framework in which to consider successful data sharing:

- **Semantics** the meaning of information
- **Syntax** the format of information
- **Quality** the confidence to re-use information
- **Rights** permission to use information
- **Trust** who is accessing information
- **Transport** how to move information
- **Information Governance** the behaviour and culture to protect and exploit information

3. This document considers each of these themes and the standards and components that could be re-used in each to support a strategic approach to data sharing.

Semantics – the meaning of information

4. To share data, either in a set or as an item, the owner will need to communicate the meaning of each value, and use consistent identifiers and vocabularies to refer to the same 'Thing'.

5. Semantics can be expressed in

A corporate data dictionary	A set of metadata maintained across the enterprise that lists and defines each data element, and where applicable, lists the codes and meanings that can be used with each element.

[1] https://knowledgehub.local.gov.uk/group/localegovernmentstandardsbody

Persistent identifiers	The business may generate Identifiers that are then re-used across the rest of the enterprise, and form links between data sets. Some identifiers are valuable beyond the enterprise, and can be used as 'connective reference data' within a National Information Infrastructure.
Logical models	A set of models maintained across the enterprise that define how elements are combined into classes and how those classes can be related. Models can be expressed as UML

6. The 'Semantic Web' provides standards that can be used to publish, and share, semantics.

Uniform Resource Identifier (URIs)	A unique string of characters, used to identify a 'Thing'. When a URI is also an HTTP URL (which they most often are), then they can be looked up to give a definition, and more information, about that 'Thing'. See http://en.wikipedia.org/wiki/URI http://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/#uniform-resource-identifier http://legsb.i-network.org.uk/resources/publishinglocal5stardata/
Ontology, such as • Resource Description Framework Schema (rdfs) • Web Ontology Language (owl)	To define properties, classes, and relationships. Each definition is available as a URI and so can be referred to individually.
Simple Knowledge Organisation System (SKOS[2])	Used to publish vocabularies, (e.g. codes and descriptions), and to link terms to other vocabularies.

7. The semantics of one department, or sector, is likely to use its own language and terms. However, the data is likely to follow common patterns that recur across the public sector, for example

a PATIENT with a CONDITION has TREATMENT at a HOSPITAL and FEELS BETTER a PUPIL learning a SUBJECT has a LESSON at a SCHOOL and gains a QUALIFICATION These could both be considered as a PERSON in a ROLE has a NEED uses a SERVICE from an ORGANISATION at a LOCATION and achieves an OUTCOME.

To make these types of joins will require an 'Upper Ontology' that gives definitions of concepts that can be used in any public sector setting, against which sector semantics can be mapped. A Public Sector Concept Model becomes a key asset when sharing data between sectors.

8. 'Semantic Interoperability[3]' can be achieved when the meaning of data is linked directly to the data itself.

[2] SKOS – Simple Knowledge Organization System - http://www.w3.org/2004/02/skos/

Data Sharing Checklist – Semantics

- Is there a corporate data dictionary that describes the meaning of each element and term?
- Have 'Persistent Identifiers' been identified and used consistently?
- Are logical models used to enforce the relationships between data items?
- Are dictionaries, vocabularies, identifiers and models published?

Syntax – the format of information

9. To share data, the owner must provide it in a format that the intended recipient can consume.

10. Where data is provided at a portal, this may be as simple as using HTML which is supported by a common browser.

11. Where data is to be published as open data, the format should itself be in an open standard, such as 'csv', for which there are choices including open source tools.

12. For some types of data, there are recognised formats, for example

    - the financial sector makes use of XBRL[4]
    - statistics suit SDMX[5].

13. Where information is exchanged to support a tightly coupled process, a specific XML schema may be developed.

14. It is useful to refer to a catalogue of data-types and reusable xml fragments when building XML schemas, such as the Government Data Standards Catalogue[6].

15. Where schema fragments have been associated with a data dictionary, XML schemas to meet a particular data sharing scenario can be quickly created.

16. Where data is published as open data, where the consumer is not known, the RDF[7] data model can be used. This model explicitly links semantics to the data itself.

Data Sharing Checklist – Syntax

- Are open formats used when publishing open data?
- Are industry standard formats used when providing data for common uses?
- Is a catalogue of XML fragments used when building XML schemas?

---

[3] http://en.wikipedia.org/wiki/Semantic_interoperability

[4] XBRL – Extensible Business Reporting Language - http://www.xbrl.org/Home/

[5] SDMX – Statistical Data and Metadata Exchange - http://sdmx.org/

[6] GDSC – Government Data Standards Catalogue – now archived at http://webarchive.nationalarchives.gov.uk/+/http://www.cabinetoffice.gov.uk/govtalk/schemasstandards/e-gif/datastandards.aspx

[7] RDF – Resource Description Framework - http://www.w3.org/RDF/

- Can open data be published as 5* data using the RDF data model?

Quality - the confidence to re-use information

17. To re-use data, the consumer must be confident that it is fit for the new purpose.

18. The data owner should therefore make statements about the quality characteristics of the data such as

- Provenance – the processes that the data has gone through, such as collection, verification, audit, aggregation and so on.
- Expectations such as accuracy, timeliness, completeness and so on

19. Sharing this information will assist a potential consumer of data to assess if it is suitable for its new purpose.

Data Sharing Checklist – Quality

- Is the provenance of data recorded and shared?
- Are data quality characteristics set, and monitored?

Rights - the permission to use information

20. For protected data to be shared, the data controller[8] needs to be assured that

> *A **person** is **empowered** by their **organisation** to act in a **role** that has a **right** to a set of **data items** for a **purpose,** and agrees to the **terms** by which the data is to be used and handled.*

21. To determine if a data share is authorised, the following information also needs to be shared

		Example
Purpose	The '*purpose*' for which data can be shared. This can be drawn from a list of public sector services and activities.	In Local Government, the esdToolkit[9] lists types of functions and services, linked to the legislation that gives the relevant powers and duties.
Data Items	The set of *'data items'* that support the '*purpose*', drawn from a data catalogue.	
Right	The legal basis for the disclosure of the set of '*data items*' being shared. This may refer to defined legal gateways.	HMRC publish all the legal gateways, by which that can share data. See http://www.hmrc.gov.uk/manuals/idgmanual/IDG50000.htm
Terms	The Information Governance undertakings necessary, or the licence terms.	A Local Government Information Governance Toolkit is being set up, based on the Department of Health Information Governance Toolkit, which lists 'measures' that

[8] http://www.ico.gov.uk/for_organisations/data_protection/the_guide/key_definitions.aspx
[9] http://standards.esd.org.uk/

		can be combined to provide a set of IG terms. See http://legsb.i-network.org.uk/promoted-standards/information-governance-toolkit/ .

22. There will be occasions when the user does not need the actual data items, but wants information that is derived from one or more data items. This may be because

- The detail of individual data items is too complex for a person who is not a professional in that discipline;
- The requirement may be for less sensitive data.

23. Examples of Derived Data may include.

Data Item	Derived Data
Date of Birth	Person is over 65 years of age.
In receipt of Jobseekers Allowance (Income Based)	On a passported benefit
Has had an epileptic fit in the past 12 months.	Unfit to drive

24. It should therefore be possible to define business rules acting on the data catalogue, to produce intermediate results. This derived data may:

- require a lower strength of assertion;
- require less controls to handle and protect;
- require less interpretation as it is used in other disciplines.

25. Open Data, has a simpler set of requirements for authorisation, which requires that the user agrees to the licence conditions. The Public Sector Transparency Board recommends the use of the Open Government Licence[10].

Data Sharing Checklist – Rights

- Are the legal gateways etc by which data can be shared, associated with each data set?
- Are data sharing facilities established against a catalogue of 'purposes'?
- Are Terms of Use drawn from a re-useable set of Information Governance measures?
- Is there a consistent licencing regime for the re-use of information?

Trust – who is accessing information?

[10] Open Government Licence (OGL2) - http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

26. For protected data to be shared, the data controller[11] needs to be assured that the person requesting access, is the same person for whom a permission has been agreed, using the Rights theme.

27. An assessment of the risk of a loss of confidentiality, leads to a consideration of the level of certainty required as to the identity of the user.

28. A credential, often contained in a token, is provided to a user via processes that are trusted by the data controller, to provide the required level of certainty.

Data Sharing Checklist – Trust

- Is there a consistent scheme to assess risk for loss of confidentiality, which leads to a determination that is shared and understood?
- Is there a trust regime that enables credentials to assert a level of certainty of identity that matches risk levels?
- Is there agreement on security processes such as 'identify verification'?

Transport - how to move information

29. The means by which information will be moved will be determined by factors including an assessment of the risk to confidentiality, availability, and integrity. That is …

Confidentiality	Does the transport mechanism protect the information form unauthorised access?
Integrity	Does the transport mechanism ensure that the information is not altered between sending and receiving?
Availability	Does the transport mechanism operate reliably at the speed and times required to meet the needs of the business process?

30. As a principle, it may be preferable to avoid making copies of data which is then sent to the recipient, as further controls and governance are then required on the copy. Where appropriate, it may be better to give access to the data 'in-situ', perhaps via a portal, or an api.

31. Many options exist as data transport mechanisms, including

- the post
- a courier
- telephone
- fax
- email
- web site
- transmission over a network
- access to an ai

… each with their own risk characteristics.

[11] http://www.ico.gov.uk/for_organisations/data_protection/the_guide/key_definitions.aspx

32. Some networks require that a 'code of connection' (CoCo) is met before an organisation is allowed to use it. CoCos(s) typically assure that the organisation's perimeter is secure, and that it has sufficient governance procedures.

Data Sharing Checklist – Transport

- Is there a consistent scheme to assess risk for loss of confidentiality, Integrity, and Availability to a business process, which leads to a determination that is shared and understood?
- Can each potential transport mechanism be mapped to risk levels?

Information Governance - the behaviour and culture to protect and exploit information

33. Each data sharing scenario typically comes with a set of undertakings that the receiving organisation accepts which define the purposes that the data may be put to, and how it will be handled through its data-lifecycle.

34. Without some coordination, or a base set of measures, there can be many of these arrangements, each with its own audit and enforcement regime. Local Authorities, who share data with many sectors, can find that they must comply with a number of these separate arrangements.

35. Information Governance arrangements are typically based on ISO27001.

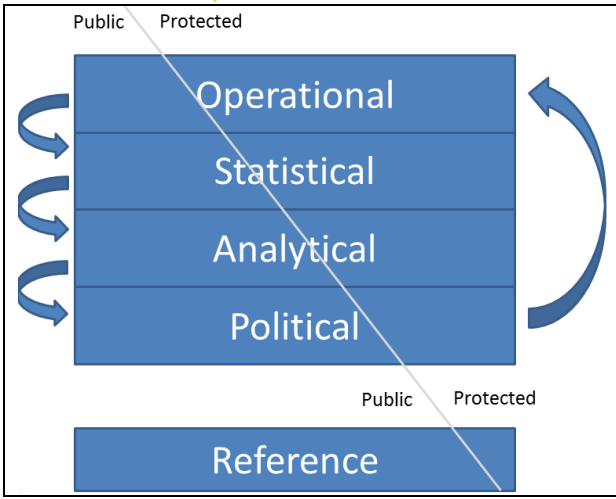
36. The Department of Health Information Government Toolkit is a leading example where a base set of measures is defined, from which Information Governance arrangements are built for each data sharing scenario. See https://www.igt.connectingforhealth.nhs.uk/

Data Sharing Checklist – Information Governance

- Is there a catalogue of base 'measures' that can be combined to meet the Information Governance requirements of each data sharing scenario?

Contexts of Data that can be shared.

37. Public Sector Data can usefully be considered over five contexts.

	**Operational** - about real people and places, with real needs and circumstances, using real services, e.g. case work; **Statistical** - aggregated operational data, organised using common classifications and segmentations; **Analytical** - the conclusions drawn from an analysis of statistical data; **Political** - the decisions taken to shape services, e.g. budgets, strategies, priorities, targets etc; **Reference** - master data to give common identifiers and definitions to objects that can be used to link data;

38. Each dataset, will have been assessed by the owner as to whether it is public, or protected. That is …

Public Data	"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." http://data.gov.uk/opendataconsultation/annex-2
Protected Data	Data … • containing personal information which Is covered by the Data Protection Act, or • for which there is a relevant exemption from legislation such as the Freedom of Information Act. or • containing IPR, which itself maybe from a 3rd party, which requires special licence conditions or fees as described by the regulations for the re-use of public sector information.

39. Data that may be protected in one context, may become public when processed into another context. For example, data about people and their circumstances may be protected, but statistics about caseloads and segmentations may be public, and valuable insight might be released.

40. Consequently, it will be useful to mark each dataset in an Inventory with a simple classification scheme to indicate if the data is considered

    Context
    - Operational
    - Statistical
    - Analytical
    - Political
    - Reference

    Openness
    - Public
    - Protected
        - Personal Data
        - FoI Exemption
        - Intellectual Property

41. It would also be useful to be able to make links across an Inventory to show where a Public dataset has already been derived from a Protected Data Set.

Building an Inventory of Datasets that could be shared

42. As a response to the Shakespeare Review[12] (reporting in 2013) , the government have asked each Department to create, and publish, an Inventory of all of their datasets (not just those that could be published as open public data).

43. Departments who follow the CESG Information Asset Maturity Model[13], will already have an Information Asset Register that lists their protected assets and details how each is controlled.

44. This is an opportunity to both

- support the drive towards publishing open data
- promote datasets that can be re-used.

45. An Inventory could list

Semantics	• The Identifier Schemes used. • The core concepts contained in the data
Syntax	• The formats that the data is available in
Quality	• Provenance • Data Quality characteristics
Rights	• Licence Terms • Legal Gateways applicable
Trust	• Risk Levels • Acceptable Trust Schemes and Credentials
Transport	• Risk Levels • Acceptable Networks
Governance	• MoU(s) • Information Governance Measures

[12] https://www.gov.uk/government/publications/shakespeare-review-of-public-sector-information
[13] CESG Information Assurance Maturity Model http://www.cesg.gov.uk/policyguidance/IAMM/Pages/index.aspx