Introducing linked open data: primer

The phrase Linked Open Data (often abbreviated to LOD) contains two relatively recently defined (or redefined) concepts: "open data" and "linked data"[1]. Let's look at these in turn:

Open Data

Many organizations hold large quantities of non-personal data – from details of projects they run or fund, to geographical information on issues of interest to them, to catalogues of research, to performance statistics and evaluation data on their programmes. In many cases, this data is kept in internal systems, and only provided to a small number of authorized people. Sometimes this data is provided to others at a cost. Sometimes this data may be made freely available on websites or in online downloads.

Much of the value in this data can only be gained when it is combined with other datasets, or used by independent actors who are able to use data to ask new questions and explore issues that the origional data owner had not considered. Making data available as open data removes the organizational, technical and legal barriers to data being re-used by third-parties. Eaves (2009) suggests three broad principles of open data which, drawing on the Open Knowledge Definition (OKF - Open Knowledge Foundation 2006), can be re-framed as stating data is open only if:

- **It is accessible and discoverable** – generally through being made available for free download online.

- **It is in open, machine-readable formats** – so that the user does not need proprietary or expensive software to explore and 'remix' or manipulate the data. There should be no technical restrictions to prevent re-use of the data.

- **It is available under an open license** – or placed into the public domain. A license should not restrict who can use the data, or place limits on what they can do with it (including redistributing it). A license might request that the source of the data is credited (attribution), but should not place other burdens on data re-users.

It might be added that data is also only effectively open is any code-lists and documentation necessary to interpret it (e.g. details of the units of measurement used etc.) is also made openly available.

Open data is published in a wide variety of formats, from Excel (XLS) and Comma Separated Value (CSV) files that can be accessed in desktop software, through to XML and JSON feeds provided in bulk downloads or through web service APIs, and specialist formats such as geographic shape files.

[1] At least in the context of this paper, and in the sense of formal definitions.

Box-out: A question of commerce

The Open Knowledge Definition (OKD) specifically includes terms stating that licenses for open data (and other open knowledge) should not discriminate against commercial re-use (some licenses such as Creative Commons Non-Commercial seek to prohibit third parties profiting from content whether other copy rights have been otherwise waived). It has been argued that the insistence of the OKD on permitting commercial re-use can be problematic in some contexts[2] (See for example Wright et al. 2011) and if an all-or-nothing approach is taken to this term it could block the growth of an open data eco-system for civic use. In the Access to Knowledge movement, the potential impact of open licenses on local and indigenous knowledge has also been explored, with concerns raised that local knowledge could be unfairly appropriated or exploited by those with established economic power, to the disadvantage of the communities who have stewarded local knowledge over generations.

However, proponents of pro-commercial OKD terms argue that (a) commercial actors have a key role to play in building the open data eco-system, and there is a risk that non-commercial licensed data will be excluded by license from important data-services and platforms; and (b) that non-commercial licenses risk introducing significant downstream license incompatibility, limiting the extent to which a non-commercial licensed dataset can be included in mash-ups or integrated with other systems that have more permissive licenses. For example, content on Wikipedia is licensed under a Creative Commons Attribution Share-Alike license. Non-commercial licensed data could not be included on or added to Wikipedia pages and info boxes[3]. There may be a tension between the needs of individual knowledge holders, the preferences of organizations, and the conditions most conducive to the development of rich and comprehensive open data resources.

Linked Data

Different datasets held by different organizations around the world might contain overlapping content. Sometimes two organizations contain similar data about different areas of the world: for example, two research libraries – one holding information on development in Asia, the other focused on development in Africa. Sometimes datasets will contain different information about the same topic, entity or place. For example, a datasets detailing where aid money has been spent and who it has been spent with, and another dataset listing outcomes produced by some of the same organizations who received money, or a information resource with stories, narratives and media

[2] For example, where particular economic exploitation of a democratic dataset could lead to citizens being harmed. Ruling such as dataset as 'non-open' potentially acts rhetorically against initiatives to release such datasets at least for civic re-use.

[3] This is a generalization. The intellectual property and copyright regime around databases varies widely across the world. See OpenDataCommons.org (Hatcher & Waelde 2007) for more details.

from citizens living where aid money was spent.

Conventionally these overlaps are hard to find, and when found, connecting different datasets together is a laborious task. One dataset might store location information on research using country codes, another might attach research records to regions. In one dataset there might fields detailing the full contact details of every organization funded; in another dataset just initials are held. The variations could be endless.

Linked data puts forward a set of technical and social standards and conventions that make it easier to connect different datasets.

Firstly, it provides a data model in which all information and data is represented using a collection of 'triples'[4], generally recorded using RDF (see box-out).

Triples are simple building blocks of data representation. Take the simple table below for example:

Meetings	**Location** (P)
IKM Linked Data Workshop (S)	Oxford (O)

We can restate the central fact asserted in this table as a simple relationship:

(1) "IKM Linked Data Workshop" took place in the "Location" of "Oxford"

The triple is made up of a subject (S), predicate (P) and object (O). You can think of the predicate as a link, we've linked the "IKM Linked Data Workshop" to "Oxford" using a "Location" sort of link. Links then are at the heart of the linked data model. Everything a linked dataset contains is build up from these building blocks.

You will notice that statement (1) doesn't contain all the information in the table. We also need to add an extra statement to say that:

(2) The "IKM Linked Data Workshop" "is a" "Meeting"

These triples are not formatted in any particular linked data standard (such as RDF) yet, but this example should point at the possibility for simple triples to represent the contents even of complex relational database.

Secondly, the linked data model encourages the use of URLs as identifiers

[4] In fact, many systems now use 'quads' by collecting up a set of triples in a 'named graph' to make it easier to pick out sets of triples without overly complex data models. However, this detail is not relevant for understanding basic linked data principles.

for things and relationships in a dataset[5].

For example, instead of using the literal string of text "IKM Linked Data Workshop" as our subject, and "Oxford" as the object, we might use web addresses to refer to these 'things'. We could also find that someone has already established a URL which defines the meaning of Location sort of links, and we could use that in place of the text "Location". This would give us a triple that looked something like:

(3) <http://events.ikmemergent.net/IKM_Linked_Data_Workshop> <http://purl.org/dc/terms/Location> <http://dbpedia.org/resource/Oxford>.

This might not be so easy for a human to read, but it's a lot more useful to a computer. Particularly if two different datasets agree on using the same URLs to identify a think. For example, we've used a URL at dbpedia.org to refer to the City of Oxford[6]. This performs at least two useful functions: (1) dbpedia.org is widely used in the linked data community to identify things, so there is a significant chance that at least some other datasets referring to Oxford will have also used this an identifier for the city; (2) If I look up http://dbpedia.org/resource/Oxford I find a clear description that tells me "Oxford is a city, and the county town of Oxfordshire, in South East England." The URL is far less ambiguous than simple the string of text "Oxford" which could, depending on the context of the reader, have been taken to refer to Oxford in Pennsylvania.

Notice that we've also used a URL for the relationship in this triple. http://purl.org/dc/terms/Location is a URL taken from the Dublin Core Meta Data Initiatives vocabulary of terms. Linked data vocabularies provide lists of terms which can be used in describing things and relationships in a dataset. You can think of them a bit like a library of column headings. There are many different vocabularies available, and anyone can create their own, or extend an existing vocabulary. When two datasets converge on the use of a vocabulary, combining them and understanding that they contain similar sorts of things and relationships becomes easier.

Thirdly, the linked data model encourages identifying things with URLs that return useful content when they are looked up (dereferenced).

Ideally if humans request content by looking up a linked data URL in a web browser they will find human-readable content, and machines that look up a URL will find data provided using established standards such as RDF and SPARQL (a query language for triples).

---

[5] Technically the model asks for URIs (Uniform Resource Identifiers) which are distinct from URLs (Uniform Resource Locators), but in practical LOD contexts we deal primarily with URLs.

[6] DBPedia is a community-driven effort to represent structured information from Wikipedia as linked data.

The information returned might include textual descriptions of the thing referred to by the URL, or it might also include other known data about it, such as it's co-ordinates, age, categorisations or other relevant facts. This additional data might use a mixture of literal strings of text (e.g. for the name and description of a thing), and URLs, allowing an agent accessing the data to navigate from link to link, picking up data from diverse distributed datasets.

Tim Berners-Lee's (2006) linked data principles (often referred to as the 'Four Rules of Linked Data'), explicitly encourage publishers of linked data to include useful links to related datasets (for example, using the rdfs:seeAlso and owl:sameAs properties) just as public spirited website owners often include a list of links to further sources of information on their web pages.

In following sections we will spend more time exploring practical technical, policy and governance issues that the linked data model gives rise to.

Box out: RDF

RDF stands for 'Resource Descriptor Framework'. RDF is a data *model* that is based around triples. Generally talk of linked data is talk of RDF, but linked data can also be published using microformats and other non-RDF approaches.

RDF can be serialized (turned into flat files of data) in a variety of formats. One of the most common is RDF-XML, which uses XML as an interchange format. However, RDF can also be serialized as n-tuples, Turtle and in N3 syntax – variations on a plain text syntax in which triples are written out.

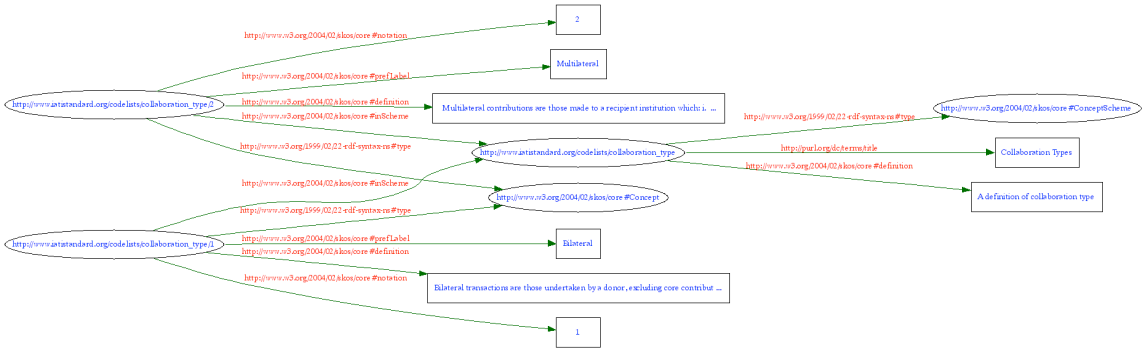
In this report I generally use N3 syntax, which provides a more human readable abbreviated syntax where subjects and predicates need not be repeated (with , and ; used to indicate objects, or predicates and objects are being added to the prior subject & predicate / subject).

There are free converters online that will convert between RDF-XML, Turtle and N3, and the *cwm* python script, documented at http://infomesh.net/2001/cwm/, can be downloaded and run on your own machine to convert between formats.

At http://linkedinfo.ikmemergent.net you will find a short animated presentation providing an alternative introduction to linked data that may be useful to visual thinkers. Visual thinkers may also find the common representation of linked data as a 'graph' useful. Whereas a common way of thinking about relational databases is as tables of data with some linking key columns, linked data is more commonly visualized as a graph of Resources (identified with URLs), predicates linking them, and string literals to label things or where a text description is being used in place of another Resource. The network-like diagrams generated are referred to as 'graphs'. The term

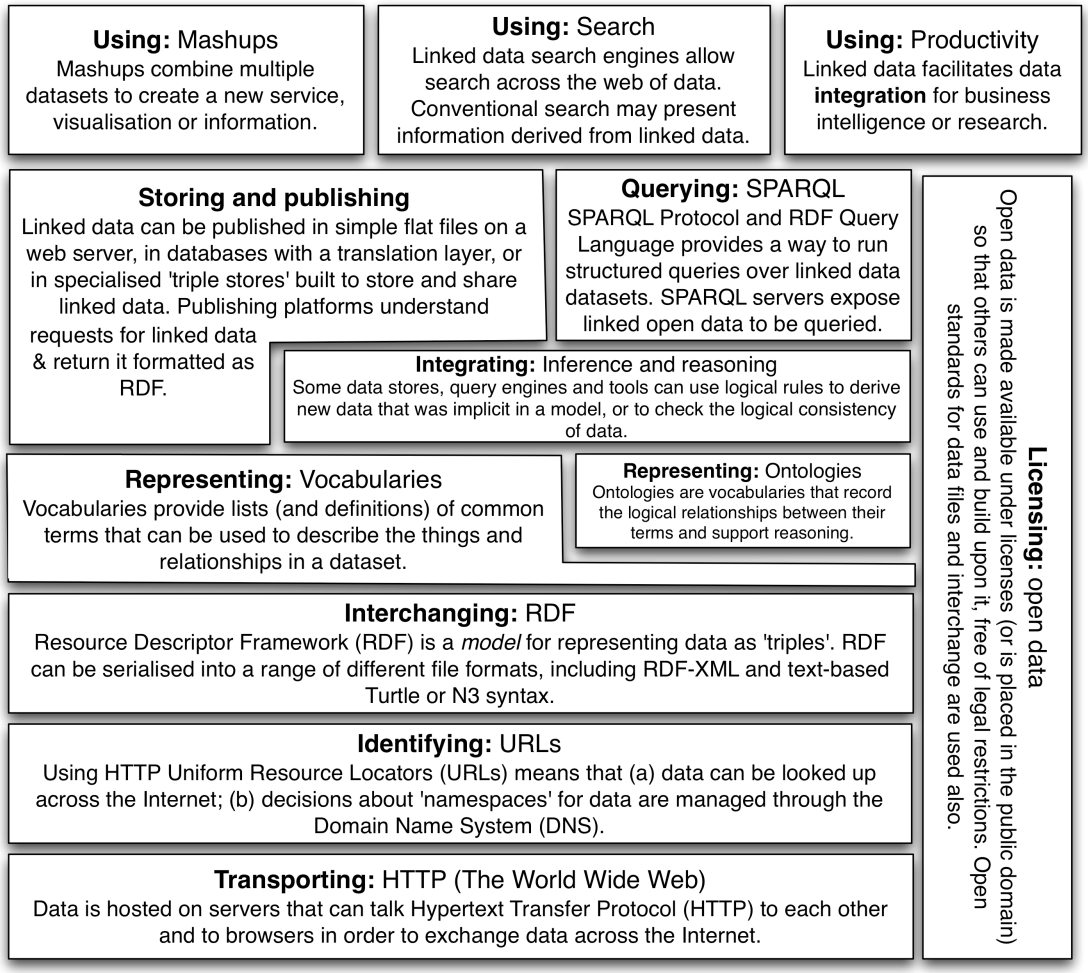
'graph' is also commonly used to refer to specific sets of linked data triples.

The graph below was generated with RDF data using the W3C validator service (http://www.w3.org/RDF/Validator/).



Linked Open Data

It is possible to have open data that is not linked, and linked data is not open, but it is when the two are combined, in linked open data, that a global web of interconnected datasets can emerge. Understanding what this web of data is and how to make international development relevant data a part of it, involves understanding a number of different jigsaw pieces.



Elements of the Linked Open Data Stack (revision 4) - 5th May 2011. CC BY-SA-NC
Draft sketch by Tim Davies (@timdavies / tim@practicalparticipation.co.uk) for IKM Working Paper on Linked Open Data for Development. Comments welcome. Search 'linked open data stack' on http://www.opendataimpacts.net for latest version.

Idea based on Semantic Web Stack at http://en.wikipedia.org/wiki/Semantic_Web_Stack

The diagram above uses the metaphor of a 'technology stack' to show the different pieces involved (if such diagrams are unfamiliar, don't worry – the most important elements will be explained in more detail in the coming sections). Reading the diagram from bottom of top you can see how different technical, organizational and social systems are involved in publishing and using linked open data. These range from the technical protocols of the World Wide Web, through to the organizational structure of the domain name system that gives control of particular web addresses (namespaces) to particular organizations and to allow them to represent things referred to using identifiers pointed at that namespace, and to the design or selection of vocabularies and ontologies (often based upon established standards within a

sector) that provide 'terms' with which to represent data using the technical RDF data model.

Not just a technology stack

Publishing data as linked open data provides the foundations for a wide range of applications, often by removing the technical, legal and organizational barriers to combining and working with multiple distributed datasets. However, the choices you make about your own linked data stack, such as which sets of URIs (identifiers) to make use of; what tools to provide re-users of your data and which tools to adopt internally; and how you consume as well as produce linked data, will impact on how potential benefits from linked data approaches are distributed.

The following section shares some brief case studies of linked data demonstrator and pilot projects in the development field, and the results of a short mapping study to explore the current scope of open development data. The report then turns to offer practical guidance, before ending with a summary of key considerations for the development of a pro-development linked open data eco-system.