Case studies: linking development data

The following case studies detail four linked data related projects. The first and third were commissioned by IKM Emergent as demonstrator projects. The second and fourth have been commissioner or carried out by participants of the Linked Open Information workshop held in Oxford in late 2010.

Young Lives Linked Data Demonstrator

Young Lives is a longitudinal study on child poverty, hosted by the University of Oxford. It is following 12,000 children over 12 years in four different countries (Peru, India, Vietnam and Ethiopia) using household surveys and child surveys, inter-household data and community data related to child health, education, employment and income, family status, and welfare to understand the causes and consequences of child poverty. Young Lives has generated significant datasets and a core aim of the research programme is to make this data more accessible to policy makers, other researchers, and practitioners. However, this is difficult to do because they are statistical datasets, currently archived using the SPSS statistical software format#. We worked with the Young Lives team to explore how converting some of the data into linked open data could support the goal of increasing data accessibility and use, whilst exploring the issues involved in creating linked data.

We identified a sub-set of the data for a trial - focusing on child-level health data from the second round of the study for a single country (Peru). We used custom scripts and existing open source software libraries to convert this data into the RDF linked data model. Using a tool called 'OntoWiki' we published this data online where it could be queried using standard linked data approaches. We also created a simple visualisation tool to compare example statistics from the dataset (e.g. comparing smoking prevalence amongst boys and girls in different regions of Peru), and converted some example comparison data of national youth smoking prevalence rates across Latin America to explore how linked data could, in theory, facilitate cross-dataset comparisons.

Generating linked data from existing tables of data is not as simple as saving a file in a new format. It involves choosing the different 'things' that will exist in the new dataset (Children? Survey responses? Countries? Regions? Questions?), and the properties that will be used to relate them (e.g. Is the answer to a question a fact about a child? Or is there a Child, who was asked a question, and the answer to that question is X?). These are data modelling choices. In Linked Data we also have to choose what properties and identifiers we will re-use from existing linked datasets.

In choosing how to model the data we adopted a number of principles:

- **Simplicity & flat structure:** RDF can allow very complex models, but these make querying the data trickier. So we sought to adopt simple structures, even if these meant less 'expressive' descriptions of the data.

- **Comparability with SPPS data:** we used identifiers that could be easily converted into SPSS variable names, so that users could check back against the 'authoritative' data.
- **Ease of annotation:** allowing additional information to be attached to questions. One of the values of the RDF model is that annotations can be added to just about anything, and those annotations can be distributed across the web of data: they don't just have to be created by the original dataset creator.
- **Re-use of other vocabularies and ontologies**: increasing the chance of comparing our data with other data, and increasing the change of existing tool-chains being able to operate against the dataset.
- **Making linkages:** allowing data from the wider web of linked data to be used in the demonstrator.

There were a number of lessons learnt from this project. First, it was important to be pragmatic about semantic modelling and finding shared vocabularies, in addition to finding the availability of linkable data. Secondly, whilst the linked data cloud is growing, there was little comparable data to the Young Lives data yet available: we had to mock-up comparison data. Whether or not linked data will facilitate easy comparison of different datasets depends both on data providers adopting technically similar ways of representing data, but also on convergence of ways in which data is collected. In the Young Lives case we were able to compare smoking prevalence data from the Young Lives study with regional data because Young Lives had drawn on existing World Health Organisation (WHO) questions in creating their survey. However, there were other data points even in the small dataset we worked with where the question asked in Young Lives was *similar to,* but not directly comparable with, questions from another study, limiting the potential for comparison. This is of course not a technical issue, but the technical possibilities of comparison may create interesting pressures for convergence of studies and definitions, with both positive and negative impacts. Thirdly, we found that the tools for converting complex datasets into linked data are not particularly user-friendly, and required high levels of technical knowledge to operate. However, the modelling process involves a high level of knowledge about the data being modelled, suggesting that linked data creation will often require (at least in the short-term) significant collaboration between domain experts and technology experts.

We also encountered issues during the project about the degree to which the survey data could be 'open'. At first, we sought to publish anonymised survey results, taking care not to name the region of child respondents unless certain anonymisation criteria were met. However, on reviewing the policy around data release, we realised we could not openly publish this information online, but needed to retain the licencing arrangements of the UK Data Archive where this detailed survey data is kept to protect survey respondents. Instead we focused on publishing the questions asked in the survey, and generating some summary statistics from the data.

A further IKM project, building on the learning from this stage of the project is currently underway to express the full question set from Round Three of the Young Lives study as linked data, as well as selected summary statistics, providing a permanent location for these online to allow them to be linked against.

The Global Hunger Index as RDF

The Global Hunger Index is published annually by the International Food Policy Research Institute (IFPRI) and partners, as a statistically generated score ranking countries between 0 (no hunger) and 100 (the worst) based on the proportion of the undernourished as a percentage of the population; the prevalence of underweight children under the age of five; and the mortality rate of children under the age of five.

In 2010 an Excel spreadsheet of the Global Hunger Index data, including the source data used to calculate it, was publised as open data by IFPRI alongside the editorialised 52-page Global Hunger Index report. This open data was used by newspapers such as The Guardian to create their own analysis of the data, and to invite their readers to explore the data. A number of bloggers and other interested parties independently created their own visualisations of the data. IFPRI published it's own custom google maps visualisation of the data alongside the report on it's website, and this was picked up by a number of parties and embedded in other websites and blogs.

To explore how linked data modelling could complement the existing Global Hunger Index open data, in late 2010 IFPRI piloted the creation of two linked data versions of the index. These were created using Google Refine, desktop software that requires minimal programming knowledge (some use had to be made of the 'Google Refine Expression Language', somewhat similar to using formulae in a spreadsheet). The two linked data version of the GHI used different 'models' to represent the data. The first, SCOVO, was simpler, but more limited in it's ability to 'annotate' GHI values; the second, the RDF Data Cube Vocabulary (QB) was at earlier stages of it's development, but allowed each GHI index value to be annotated with notes on the accuracy of the data, or any estimates that went into creating a particular countries index value due to missing data. All these annotations were originally included in the Excel open data using color coding or annotations designed for human readers rather than for computers to read. Website statistics indicate that the second format (Data Cube) has been far more widely used than the SCOVO modelling of the data.

There are many different ways to publish linked data on the web. To get the benefit of the query language for linked data (SPARQL), special linked data publishing tools are required. However, small sets of linked data can also be published directly on a standard web server. This was the approach adopted for the Global Hunger Index, which, as a small dataset, was made available at http://data.ifpri.org/rdf/ghi/. Although the skills required to work directly with linked data remain rare (and initiatives like the UK Governments Data.gov.uk who were early adopters of large-scale linked data publishing have been

building simpler interfaces onto their linked data that export it in more developer formats like JSON and XML, and common spreadsheet formats like CSV), the linked data version of the GHI did support it's integration into the Food and Agriculture Organisation's (FAO) country information portal (REF), bringing GHI information to a far wider audience.

The Global Hunger Index pilot demonstrated that for small datasets, the barriers to publishing linked data are falling as simple publishing tools become available, and where the 'linking point' between datasets is something like a country, connections between datasets are easier to envisage and make. Even so, the conversion of the data involved making choices about which identifiers to use for countries - as initially we made use of the GeoNames.org service, rather than the FAOs own linked data identifiers for countries - which potentially introduced some complexity for the FAO to reconcile these identifiers to their own. Selecting a set of identifiers for concepts like countries in a dataset can be both a pragmatic, and a political, decision.

The experience of IFPRI with the Global Hunger Index also shows the potential of open data for advocacy: releasing compelling data behind reports can increase the attention that the issue and the report gets, and can lengthen the life-cycle of important research work.

IKM Vines - Linked Information

IKM Vines is a demonstrator project exploring ways to combine information from different sources and to surface content from the South more prominently than it tends to be in conventional search engine results. Vines reads in textual information from articles tagged using the delicious social bookmarking platform, or shared in RSS feeds, and it then uses 'tag extraction' tools to find additional tags and key words relating to content. The tagged information, which, in the Euforic Vines prototype can include video as well as text, is made available to search and browse. For navigation, the interface presents 'leaves' on the left of the screen for the most significant categories of data uncovered, with a search box and tag-cloud on the left to help visualise the most prominent topics in the currently displayed content. Vines seeks to support both the discovery of information from the South, and enable an exploration of the particular sets of terms different communities use to discuss a subject area.

Linked data facilitates the creation of relationships between the tags used in Vines. Using data from Codesria, a partner organisation from the South, the Vines team employed terminology from a classic thesaurus. They explored adding structure to the Vines list of key words through hierarchical organisation of terms from broader to narrower; and through more complex relationships. For example, a thesaurus could record that:

> *The **bee** is an **insect**.*

So that searches for information on insects also returns information on bees. There is a clear hierarchical relationship here between the word *bee* and

insect. We also know that there is some sort of relationship between *bee* and *honey*. But this is not a hierarchical relation. In this case, some sort of qualifier is needed to define the relationship between the two words.

Eg. *The **bee** produces **honey**.*

The relationship between the bee and honey now becomes more evident. This articulation of the semantic relationships between terms is often called *ontology.* In the agricultural sector, old fashion thesaurus are often still used (often having been developed for library systems), but linked data can allow a far richer range of relationships to be recorded, modelled and used to retrieving information. However, to use these relationships a system needs to be aware of them, and the Vines project has surfaced a number of questions:

- Can we make increased use of ontologies in the future to organise development knowledge?
- Can organisations rely on them?
- Who is going to use them?
- How will they be maintained?

The Vines project has also recently turned to explore the need to map together different vocabularies, thesaurus and ontologies: connecting up legacy systems for organisating information, and bridging between different fields which have their own terminology and sets of linked data identifiers for concepts and things. For example, Vines can tag content both using the commercial Thomson-Reuters 'Open Calais' system, and custom term lists of development-specific terms. Using only Thomson-Reuters Open Calais may miss key development-specific terms and lead to key information being hidden. Using only a development-focussed list of terms may restrict the connections that can be made between fields. Mapping different vocabularies together can be a complex and expensive task, and often when organisations do this they do it internally and for their own purposes, not sharing the results. Creating tools to visualise the connection between different vocabularies, and to facilitate easy mapping between them by non-technical users is the next focus of the Vines project.

Vines adds meta-data to information, and in the process makes it possible to discover, sift and sort the information in different ways. However, it does not reduce information to data: you can always get back to the original articles and media, and even follow links to access them in their original context. The particular tagging services and vocabularies an instance of Vines chooses to use, and the sorts of queries that the interface chooses to make easier, will affect which knowledge is surfaced by the tool. Tools like Vines can only make use of mappings between vocabularies that are released as open data, and it is far easier to use them when they adopt standard linked data models like SKOS (Simple Knowledge Organising System) to represent the mapping.

FAO Linked Data

The Food and Agriculture Organisation (FAO) have been actively exploring the use of linked and open data over recent years. This has included making the AGROVOC multilingual structured thesaurus, first developed in the 1980s, available as linked data; publishing country profiles and identifiers as linked data; and creating tools that support knowledge creators to take their information against these terms.

AGROVOC a traditional thesaurus, with about 30,000 concepts, that has produced 600,000 levels, in roughly 20 languages. It is a concept-based thesaurus with ontological-based relations. Whilst Excel output has been one of the main options available in the past, AGROVOC terms are now published also as linked data, with a URI for each term in the thesaurus. Mapping has also been carried out between AGROVOC and the European Union's EUROVOC thesaurus using a linked data model, so it becomes possible to search across information tagged in either AGROVOC or EUROVOC. In this case, the FAO have taken on the task of carrying out the mapping (no small task). Mapping two vocabularies often involves careful judgements, so users have to choose whether to follow the links that AGROVOC includes, or to create their own mappings. This highlights the choice between the cost of carrying out independent mapping; or the requirement to trust the decisions made by third parties in the links they create. In our IKM Workshop this was summed up by the phrase "The economics of integration; vs the politics of delegation".

The FAO have developed a range of other products to support their information management, including FAO Authority Lists for search across journals and other resources. An authority lists based approach to knowledge management, where the controlled lists define who can make assertions about what, contrast with the 'Anyone can say Anything about Anything' (AAA) principle of the web of linked data (REF - SEMANTIC WEB WORKING ONTOLOGIST), but highlight that within the wider web of data, trusted authorities may remain significant players shaping choices made about which linked information sources to use and trust.

The FAO have not stopped at making their open data available. They have also worked to create tools that use it - both as hosted services, and by developing and releasing open source software which other people can adopt.

First there is the AGRIS search engine, a database maintained by FAO including 2.6 agricultural information million records from all over the world. It is free and accessible online. AGRIS can link with indexed material in Eurovoc and takes advantages of links and relationships in the thesaurus model to help users discover relevant content, even across languages and discovering related articles that using different terminology from those the user may have adopted in their search.

There is also AgroTagger, a prototype service that uses AGROVOC to tag unstructured text with key terms. The service is slightly less advanced than commercial offerings like the aforementioned Open Calais which uses

advanced 'Natural Language Processing' (NLP) to extract terms from text, but in focussing specifically on agricultural information, an area where Open Calais is weak, AgroTagger can offer more relevant results, and returns tags that could potentially be connected into the AGROVOC web of linked data.

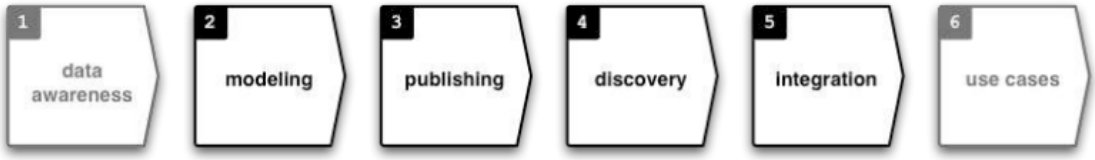
Lastly, FAO have been contributing modules to the Drupal open source content management system to make it easier to use AGROVOC tags from within Drupal, helping diverse information providers to harmonise their categorisation of content, as well as releasing as open source the AGROVOC workbench tool used to maintain the full AGROVOC vocabulary.

FAO experience demonstates some of the potential of linked data to connect up disparate information sources, but also the complexity and cost involved in doing this effectively. It also highlights the value of combining open data with open source - to make it easier for diverse local groups to participate in knowledge creation and tagging, and it highlights the importance of considering the governance of key sources of authority and linkage in the web of linked data.

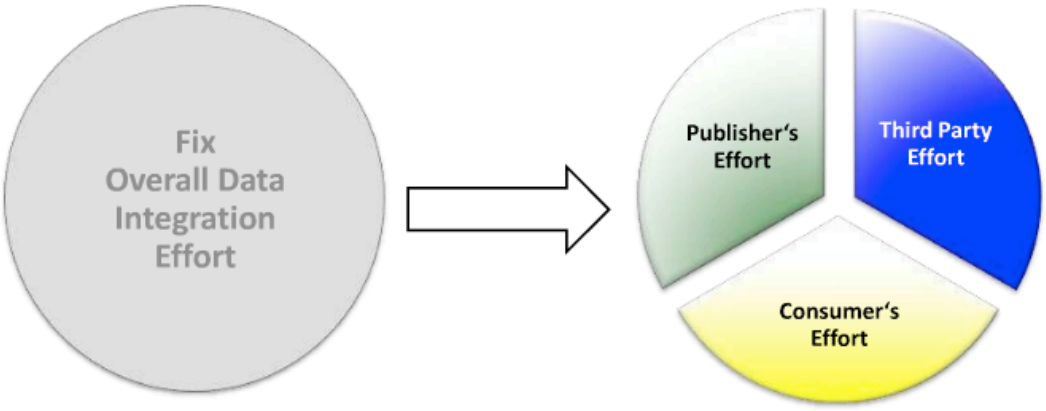
Key learning

One of the overriding lessons from our demonstrator projects is that creating linked data takes considerable investment of time and effort, and that it is not just a technical process. The return on investment in linked open data publishing may depend on who else is generating linked data at present (and thus, what connections the data can make), although there may also be some strategic benefits to being an early mover as a publisher of linked data.

The Linked Data Life Cycles project at http://linked-data-life-cycles.info/ offer a useful summary of learning from the EU funded Linked Data Around The Clock (LATC) project that matches with our findings. They describe the many stages involved in a linked data projects, from discovering data to publish (and to link to), modeling data, publishing it, promoting it's discovery, integrating data, and developing resources to support it's effective use .



They describe linked data as supporting 'Pay as you go' integration, where the overall effort of connecting datasets is shared between publishers, third parties and consumers of data. This highlights that short-term pilots with linked data may not see the full potential, as much of the value can come in the future as particular users choose to work with a dataset and are able to contribute back their integration efforts, rather than integration work being private and frequently duplicated.

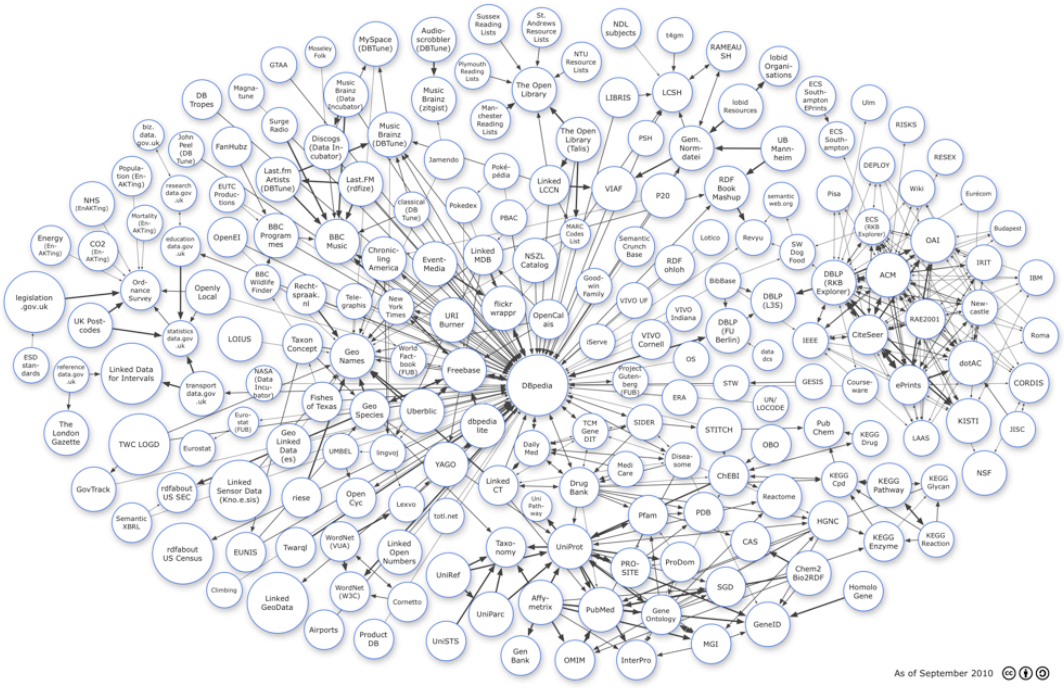


(Diagrams from http://linked-data-life-cycles.info/. Maintained by Michael Hausenblas and Richard Cyganiak)

This highlights one direction for future research: our IKM pilots have primarily focused on the technical process (and benefits) of organizations generating linked datasets from their existing data. However, many benefits may also arise from addressing the culture and skill base inside organizations (for example, amongst analysts and researchers) so that when they are working with third-party data, this also involves the use of linked open data tools, and the contribution of efforts to integrate and work with data back to an open data eco-system.

Understanding the potential of that eco-system requires some understanding of a current baseline. The following section reports the results of a preliminary mapping exercise looking at the available of open data, and linked open data, relevant to the development field.

Mapping linked & open data in development



A brief look at a recent version of the Linking Open Data cloud diagram (Cyganiak & Jentzsch 2010) doesn't show the presence of many linked data datasets that stand out as specifically relevant to international development. In June 2011 the 'International Development' category of the CKAN.net open data catalogue listed just 22 datasets. This appears at first glance to indicate that there is limited open data, let alone linked data, available in the development field. Yet assessing the state of the 'development' open and linked data web in is not a straightforward task. International development is a broad field, encompassing many areas of activity, and there is no neat boundary to divide development data from other datasets. Data about a governments spending on education is, to the country concerned, government data, although a donor agency may consider it to be 'development data'. Datasets on trade, innovation, economics and environment could all prove to be useful to development actors, though how grassroots development actors and international agencies understand the developmental relevance of particular datasets could vary significantly.

Given these complexities, in this paper offers some preliminary remarks on linked and open data in the broad domain of development based on a five-day research project. To assist an assessment of the field, we've engaged researchers in the UK and Angola/South Africa to carry out brief desk research exercises, identifying globally and locally relevant datasets with some bearing on the field of development, broadly defined as encompassing issues associated with foreign aid, governance, healthcare, education, poverty reduction, gender equality, disaster preparedness, infrastructure,

economics, human rights and the environment (Wikipedia 2011)[1]. Each researcher was asked to compile a list of datasets or websites hosting development data, and to provide reflections on the availability of data. The resulting list of datasets and sources is the product of brief desk research, and is not intended to be comprehensive, nor to provide a statistical sample. Rather, it is intended to provide examples and 'food for thought' in reflecting on the potential of linked data in development. The list of datasets is available in spreadsheet form at http://bit.ly/9HciEv - and is likely to be refined and developed further as this research proceeds.

The list of datasets has been analyzed to provide a general impression of the types of data available, the extent to which data is provided in machine-readable and open formats, and the existence of, or potential for, linkages between different datasets. In addition, the author's own experience exploring datasets related to a series of development projects is drawn upon to provide worked examples of where linked data approaches may have a role to play in the development field.

Desk research: datasets and data sources

The full list of datasets and sources of data surfaced by our desk research is shown below. The spreadsheet version at http://bit.ly/9HciEve includes links to web pages where they can be explored further.

AcaWiki	Millennium Development Goals Indicators	European Commission Database: women & men in decision making
AidData	OECD International Development Statistics (IDS) on aid and other resource flows	The IDP database (Internal Displacement Monitoring Centre)
AidFlows	Penn World Table 6.2	Kabissa Organisation Directory
Appropedia	Reegle - Renewable Energy Information	IFPRI (International Food Pollicy Research Institute Stats)
Asian Development Bank (ADB) - Statistical Database System (SDBS)	Socio Economic Database for Latin America and Caribbean (SEDLAC)	International Development Reseach Centre Digital Library
Bureau for Research in Economic Analysis of Development (BREAD)	Statcompiler	Environment.co.za (South African Environment Ministry)
Caribbean NGO Database	Stockholm International Peace Research Institute (SIPRI) Databases	Southern African Innovation Network

[1] Taking definitions from contested Wikipedia pages may not be the most robust policy in general, but in this case where we were searching for a broad-based definition of development, the Wikipedia entry proved to offer the most inclusive available set of concepts as a starting point for our enquiry.

CCPR's Survey Database	The Rural Income Generating Activities (Riga) Database	PRODDER - South African NGO Directory
Centre for International Development at Harvard dataverse	Trade Capacity Building (TCB) Database	FINSCOPE - Financial Service Needs dataset
CIA World Factbook	U.S. Official Development Assistance Database,	The statistical clearing house
data.worldbank	UCLA Social Science Data Archive	Statistics South Africa
DataGov	UNCTAD Stat	Water Affairs
Department for International Development (DFID) projects database	UNData	UNICEF South Africa Development Indicators
EMDAT - The International Emergency Disasters Database	UNICEF - Information by country	CSIR Information Services (CSIRIS)
EU Regional Policy Inforegio - Development programmes	USAID Overseas Loans and Grants, Obligations and Loan Authorizations (The Greenbook)	SADA - South African Data Archive
EuropeAid - Beneficiaries	WANGO directory of all NGOs	National Research Foundation
FAO Stats	WHO Global Health Atlas	South African Biodiversity Information Facility
Financial Access Indicators	wikiprogress.stat	SABINET - Aggregator of Journals
Global Development Network Growth Database	World Bank Doing Business	HSRC - South African Human Sciences Research Council
Global Education Database - UNESCO Data and DHS Data	World Food Programme - Food Aid Information System	South African Wetlands Conservation Programme
Global Health Observatory Data Repository	World Income Inequality Database V2.0c	AGIS - Agricultural Geo-Referenced Information System
Health, Nutrition, Population Stats (HNPStats)	World Trade Organisation (WTO) statistics database	CADRE - Centre for AIDS Development, Research and Evaluation
Human Development Report	World Water.org Dataset	ISS - Institute for Security Studies (Africa)
ILO Labor Statistics Database	International Federation of Red Cross and Red Crescent Societies Database	NARS - National Archives and Record Service
Integrated Public Use Microdata Series International	WHO Global Database on Child Growth and Malnutrition	SAHRC - South African Human Rights Commission
Inter-University Consortium for Political and Social	The Database of International Statistical Activities (DISA)	SAMRC - South African Medical Research Council

Research Database		
International Food Policy Research Institute - Project Datasets	Childinfo	Commission for Gender Equality
International Information System for the Agricultural Sciences and Technology (Agris)	NCDC Worldwide Weather and Climate Events database	DPRU - Development Policy Research Unit (Cape Town)
Joint External Debt Hub	Geodata Portal	NGI
Mapping for Results	United Nations Global Migration Database	
Migration Policy Institute Data Hub	OECDstat Extract International Migration Database	

Observations on availability and openness of data
It is evident from the above that, whilst open data catalogues may show a limited presence of development-relevant data, there is a significant amount of development-related data accessible online.

How open is the data[2]
Of the 91 sources of data explored, just under 40% clearly provided 'raw data' accessible through downloads or via APIs. Most commonly data was presented in PDFs (not raw data) or directly in text on web pages. Excel was the most common format for downloadable data, with around 30 datasets providing either Excel (more common) or CSV (less common) for download.

At least 25 of the data sources explicitly grant permission for their data or information to be re-used, with slightly more explicitly prohibiting free re-use of the data through copyright, or by terms and conditions. A number of datasets require permission to be sought before data is re-used, and at least 5 specifically include terms allowing non-commercial re-use without written permission, but prohibiting commercial use of the data. A number of data sources consist of data aggregated from other providers, and asked re-users to seek permission from each upstream provider of data before re-using it. Finding the licensing terms was difficult for many of the data sources. A small number of data sources required registration to access data.

Linked data was virtually non-existent. The Food and Agriculture Organization geo-political ontology was the only clear linked data dataset included in our sample.

Who is providing data
A wide variety of actors are behind the datasets we identified. Governments,

[2] Preliminary analysis carried out using Google Refine.

government agencies and international institutions (e.g. World Bank, WTO, WHO, ILO, UN, UNICEF etc.) were very visible as data providers and often had the most developed data hubs, with large statistical platforms bringing together and presenting datasets from across the institutions. There was significant variation between the technical and legal openness of these different data sources, ranging from sites like http://data.worldbank.org which actively promotes open access to it's data, providing APIs for developers to integrate the data into their own applications; to sites that emphasize their own query interfaces as the only way to get at the data, providing limited opportunities to download data. Second amongst our data providers we identified a large number of academic institutions, either providing specific research data for download, or offering access to catalogues of research and research data. Again there was significant variation between individual research projects, who may be providing a legacy research database online in a custom format, or behind a custom interface and potentially no-longer updated now that the research funding has ended, and large institutional repositories – using a variety of bespoke and commonly available platforms (such as the Dataverse software from thedata.org). Although all the 'academic' data sources we identified held data relevant to a diverse audience of development actors, they rarely focus on making data available to non-academic audiences.

A small number of the datasets we identified were the outcome of specific development funded projects, with a number of biodiversity and environmental datasets resulting from development programs, and the FINSCOPE Financial Services data originally produced through DFID funding, although now sustained through a commercial model that means only headline data is openly available in PDF reports. NGO and 'open source' style community projects also provided some of the datasets we located – with Wiki-based directories of research and appropriate technology included in our list.

Given our study was based on desk-research and web-searches to locate data, and given that many of the datasets we did find have very limited meta-data available on them, and few are presently listed in data directories, we have inevitably missed many datasets. It is also quite likely that many of these will be smaller datasets, harder to locate with a web-search based method. We have also predominantly located datasets in English: there may be significant available data inaccessible in English that our research has missed. However, we note that the data we have identified covers far broader ground, and is generated by a far wider range of actors than might be involved in simple 'open government data' initiatives (Alonso et al. 2011b), highlighting that open development data and open government data are far from synonymous.

Semi-structured data and information:
A number of the data sources included in our study may not qualify formally as datasets: consisting of structured and semi-structured directories of information help on pages of a website or blog. Unlike a spreadsheet-type dataset, this data could not be immediately downloaded in a single file, sorted,

or manipulated. However, in terms of content, sites like the Caribbean NGO database (managed through pages on a blog) and the Kabissa African Organisation database (managed through a web-based Contact Relationship Management (CRM) database), are broadly equivalent. The line between 'information' and 'data' is rarely a clearly defined one. Much developmentally relevant content may exist as loosely structured information, which, through the use of different technologies such as screen scraping[3], might equally be made part of the web-of-data. In fact, certain linked data technologies such as micro-formats and RDFa allow web pages themselves to become linkable sources of data, and the RDF linked data model is not restricted to only being used to mark-up large datasets – but can also be used to mark-up and provide some data-structure around small sets of information distributed across the web.

Clusters and kinds of data

There are many dimensions along which the data sources we identified could be categorized. Our initial attempts to adopt a pre-defined thematic categorization of datasets[4] (foreign aid, governance, healthcare, education, poverty reduction, gender equality, disaster preparedness, infrastructure, economics, human rights and the environment) in order to allow some summary statistics to be presented quickly hit limitations as significant sub-themes (e.g. biodiversity; migration; children) emerged without fitting directly into the categories used to seed the research. Cross-cutting data sources also emerged, with datasets such as NGO directories including information of relevance across a wide range of themes.

An alternative to thematic classification is to try and classify data sources by the 'type' of data they include. Initial analysis suggests nine main categories, listed below with some brief observations on common characteristics of these types of data:

- **Statistical data** – covering whole populations. For example, census data, national statistics and global statistics. Often available as time-series over a number of years, and with comparisons possible between countries.

- **Academic research data** – specific datasets created for research. Often one-off samples covering a specific local area. Some time series. Prepared and maintained for academic use, or archived on completion of projects.

[3] Tools like ScraperWiki.com and Needlebase.com provide means for users to take online information sources and 'scrape' or extract structured data from them when the information is published in a generally uniform way, generating a datasets available to browse and download.

[4] The short timescale and limited resources for this study mean that grounded research approaches, with iterative analysis of the datasets (and work to address gaps in the list of datasets), has not been possible – although we would hope to be able to pursue such approaches in future.

- **Market research data** – economic statistics and research. Snapshots or time-series.

- **Policy monitoring data** – collected by specific institutions to report against targets or objectives. Often part of annual reports.

- **Organisation and project directories** – details of NGOs or other agencies working in a particular area. Generally including contact details and some scheme of categorization of focus or activity.

- **Funding data** – details of aid flows or project funding, on an aggregate or individual project level.

- **Meta-data –** usually of research reports, journals or academic papers. Providing ability to search for information resources.

- **Geodata** – maps, data on geographic infrastructure, and thematic maps on environmental or economic issues.

- **Real-time data** – including weather data and other information required in a timely ongoing fashion.

Many other ways of clustering datasets are, of course, possible. One approach that may prove instructive for exploring the potential of linked and open data is to take a particular policy or practice issue, and to ask about the different datasets that might cluster around that policy issue. For example, making decisions about agricultural projects may draw upon meta-data about agricultural research, geodata on soil types and climate in a particular area, academic research data related to that area or to specific soil-types, crops or climate, and organizational data to locate actors with expertise in a particular issue.

Making connections

The fact that only a minority of the datasets uncovered in our research are available both as raw data *and* under open licenses means that there are significant practical or legal limitations on how far linkages can be made between different datasets at present: it is tricky to merge a restricted or closed dataset with an open dataset, and to keep the results open, without significant logistical or legal challenges. Data continues to be locked inside organizational silos, both intentionally, with some organizations actively writing terms to restrict ‘competitors’ or other parties using their data, and unintentionally, as organizations choose poor formats and ways of presenting their data to the world.

However, leaving the legal limitations aside, we can ask a more theoretical question about the potential of linked data in development: where are the linking points.

If you look at the Linked Open Data Cloud diagram you find that dbpedia.org, generated by converting Wikipedia pages into linked data identifiers, acts as a hub for many linked datasets. However, in different domains there are also local 'hubs' – providing sets of identifiers that different datasets converge around and thus become possible to query across[5] and connect together. What are the potential hubs for development data? And even if they are not using the same identifiers at present, are there common classification schemes adopted across the datasets we have identified that would enable them to start linking against each other, or against common hubs in future? A full exploration of these questions has not yet been possible, though a number of remarks can be made:

- **Geography -** Many of the data sources explored contained data about specific geographical areas. Whilst the one fully linked data source identified in our mapping study (the FAO Geopolitical Ontology) does highlight relationships between different codes for identifying countries, and provides web services to convert from identifiers such as country names or ISO codes to authoritative FAO URIs[6], it (a) only covers countries and regions, and (b) it doesn't provide URIs for different identifiers for countries – only the information to allow local datasets to adopt the FAO identifiers. Finding shared identifiers for second-level administrative areas in countries (e.g. counties/provinces) or other geographical areas is difficult. The geonames.org service provides some data, but the provenance and reliability of the data is unclear, and locating the correct identifiers to use from geo-names can be complicated.

  Whilst some countries have detailed identifier sets for local geographic areas available (the UK[7], Spain[8] and Brazil[9] for example), localities in many developing countries are hard to refer to on the linked data web.

- **Projects and organizations -** Many datasets would also benefit from being able to link against 'projects' and 'organisations' (for example, linking aid flows to the projects receiving them, or research documents to the organizations who can support in the implementation of evidence-based programs). Whilst we found a number of datasets providing project directories, these varied in terms of the structure of data they held, and their coverage – and were fragmented with potential duplication between them.

  Projects like OpenCorporates.com, OpenCharities.org are providing

[5] Although as we shall see in Section 3, the technical requirements to be able to query across datasets can be considerable; links between datasets are necessary but not sufficient.
[6] http://www.fao.org/countryprofiles/webservices.asp?lang=en
[7] http://data.ordnancesurvey.co.uk/.html
[8] http://geo.linkeddata.es/web/guest
[9] http://api.comprasnet.gov.br/sicaf/doc/

linked data identifiers for corporations and charity organizations using official data sources, but (a) in many cases the reality of organizational structures (and the presence of non-constituted organizations) is not reflected in official databases, and (b) coverage of these services is limited.

- **Concepts and taxonomies** – many datasets about similar topics used completely different taxonomies to categorize their data and content. Finding shared taxonomies, or mapping taxonomies together, is important to make linkages between related data held by different organizations and agencies. Whilst the FAO's AgroVoc project gives one example of a large taxonomy creation of mapping project, we didn't identify any other such projects in our mapping, not did we identify significant 'delegation' where local datasets were choosing to use a third-party set of identifiers apart from where these came from their organizational parents (eg. UN datasets use UN identifiers for countries).

Gaining benefits from linked data in development may involve focusing on developing points of linkage as much as focusing on making existing data and content available using linked data standards.

Reflections and focus for further work

This preliminary study was motivated by a desire to gain a better sense of what the development data environment might contain, and how this might have a bearing on discussions of linked data. A number of threads of research have been under-developed or omitted from this initial study. In particular, questions of language and open data, and how the open data for development landscape looks in countries without English as an official or widely used language, and analysis taking a deeper look at the connections between small clusters of datasets, would be worthy of greater exploration.

During this study we have also come to question some of the starting assumptions about the primacy of 'datasets' as a focus of enquiry. If our concern is about impact on development practice, then a broader look at development information, taking in any digital resources which could have linked data structure attached to their directly or as meta-data, may avoid an unhelpful dataset-centric bias, with it's implicit prioritization of abstracted and codified (and frequently numerical) content over other forms of knowledge.

In conclusion

The development sector is potentially rich in data: but much of the data available is neither explicitly open nor available in structured, standardised or linked data forms. The data that is available only reflects some parts of the development community. The diversity of sectors overlapping with development, and the different goals of these sectors, makes that developing open data as a resource to support development may involve widespread engagement with different communities. Development funded projects are not being consistently encouraged to share any data coming out of their work,

and some have sustainability strategies based around selling data. Whilst research projects and short-term initiatives can add useful resources and points of connection to the open data ecology, there is an ongoing need for action to curate and connect data, and to develop the capacity of actors at both international, national and grassroots levels to create, share and make use of, good quality open data. If the development sector is to build a stronger open data commons, then significant capacity building effort is likely to be needed.