

Unlocking the potential of aid information

Jonathan Gray, Jordan Hatcher, Becky Hogge, Simon Parrish, Rufus Pollock



September 2009

Licensed under a Creative Commons Attribution License v3.0
(All Jurisdictions)

TABLE OF CONTENTS

Introduction	4
Summary of recommendations.....	5
1. Making aid information legally open	6
1.1 - Legal rights in aid information	6
1.1.1 - Why should aid information be made open?	6
1.1.2 - Aid information is protected by different legal rights	6
1.1.3 - Examples of legal rights in aid information	6
1.2 - Legal mechanisms for allowing material to be re-used and re-distributed	7
1.2.1 - By default re-use requires permission	7
1.2.2 - Using a license or legal tool to facilitate re-use.....	8
1.2.3 - 'Bespoke' and 'generic' licenses and legal tools.....	8
1.2.4 - Different kinds of sharing: from accessibility to openness.....	9
1.2.5 - Making content and data open	9
1.3 - Other law and policy pertaining to sharing aid information	11
1.3.1 - Law and policy related to sharing core government information	11
1.3.2 - Freedom of Information (FOI) laws.....	12
1.3.3 - Privacy laws	12
2. Making aid information technologically open	13
2.1 - File formats.....	13
2.1.1 - Proprietary, non-proprietary and open file formats	13
2.1.2 - Machine readability	13
2.1.3 - Formats for the semantic web	15
2.1.4 - Which formats?.....	15
2.2 - Data schemas for aid information	16
2.2.1 - Standard and non-standard schemas	16
2.2.2 - Existing schemas and standards.....	16
2.2.3 - Which schemas?	17
2.3 - Accessing aid information	17
2.3.1 - Access via an API	17
2.3.2 - Access to raw data	17
3. Making aid information easy to find	19
3.1 - Publishing on the web	19
3.2 - A registry of aid information	19
3.2.1 - Metadata for datasets	19
3.2.2 - Ontologies and tags: user driven resource description	20
3.2.3 - Linking, mirroring and storage.....	21
3.2.4 - Making the registry open	21
3.3 - A decentralised model for aid information	21
4. Appendix	23
4.1 - Legal rights in content and data.....	23

4.1.1 - Copyright.....	23
4.1.2 - Rights in data.....	23
4.2 - Existing examples of aid information.....	24
4.2.1 - World Bank Data & Statistics	25
4.2.2 - Asian Development Bank.....	26
4.2.3 - U.S. Agency for International Development - Economic Analysis and Data Services (EADS)	26
4.2.4 - World Food Programme - Food Aid Information System	27
4.2.5 - European Commission Humanitarian Aid (ECHO) - Funding.....	28
4.2.6 Organisation for Economic Cooperation and Development (OECD) Development Assistance Committee (DAC)	29
4.3 - Existing Work.....	30
4.3.1 - Legal and licensing standards	30
4.3.2 - Format standards	30
4.3.3 - Data and metadata standards	30
4.3.3.1 - Resource description	30
4.3.3.2 - Data models	31
4.3.4 - Other relevant work	31

Introduction

The International Aid Transparency Initiative (IATI) is a multi-stakeholder initiative bringing together bilateral and multilateral donors, partner countries and civil society to enhance aid effectiveness by improving transparency and making information about aid more available and more accessible. At the heart of the IATI proposal is the development of an aid information standard, including an agreement that donors, and other actors, will publish information about what they are funding in a consistent way, in an open format.

In the past few years there has been growing recognition of the value of 'openness' - in relation to software, content and data. In recent months there has been a spike of interest in the last of these, open data. In the US the Obama administration has placed great emphasis on openness in government - creating a new 'data.gov' portal to allow citizens to access and re-use official data. In the UK Gordon Brown has asked Tim Berners-Lee, inventor of the web, to help open up UK government data.

Why does this matter? The recent spate of interest in openness has not only been about accessibility, but also about re-usability. New internet technologies allow NGOs, companies and interested citizens to make compelling web applications and services based on official data. By making their data open, public agencies allow citizens to re-use official data in new ways - from generating interactive visual representations, to creating better services based on the information, to connecting more datasets together so we can ask and answer more complex questions.

Similarly in the context of international development, sharing aid information more effectively will ultimately enable stakeholders to build up a richer picture - by allowing more information to be aggregated and by allowing innovation in the way this information is represented and queried.

This paper looks at what publishers of information relevant to international development (NGOs, government agencies, research bodies, ...) can do to make their knowledge assets open. In particular it focuses on legal and technical measures that publishers can take in order to make it easy for others to re-use the material they publish. It also makes a series of recommendations towards the development of a standard for publishing aid information.

Summary of recommendations

- **Recommendation 1 - Aid information should be legally open.** The standard should require a core set of standard licenses for publishing aid information under. It should require that either:
 - (i) information is published under one of a small number of recommended options:
 - Licenses for content: Creative Commons Attribution or Attribution Sharealike license
 - Legal tools for data: Open Data Commons Public Domain Dedication and License (PDDL), Open Data Commons Open Database License (ODbL) or Creative Commons CCo
 - or that (ii) information is published using a license/legal tool that is compliant with a standard such as the *Open Knowledge Definition*.
- **Recommendation 2 - Aid information should be technically open.** The standard should require that raw data is made available in bulk (not just via an API or web interface) with any relevant schema information and either:
 - (i) in one of a small number of recommended formats:
 - Text: HTML, ODF, TXT, XML
 - Data: CSV, XML, RDF/XML
 - or (ii) in a format:
 - (a) which is machine readable and
 - (b) for which the specification is publicly and freely available and usable
- **Recommendation 3 - Aid information should be easily findable.** The standard should require that aid organisations add their knowledge assets to a registry with some basic metadata describing the information.

1. Making aid information legally open

1.1 - Legal rights in aid information

1.1.1 - Why should aid information be made open?

In most cases aid information is protected by various legal rights which restrict what others can do with it. Permission must be given to allow others to reuse aid information, whether in new kinds of web applications, in interactive visual representations such as Gapminder, or in recombination with other data sources.¹

Hence, if publishers of aid information wish to encourage innovative reuse they should use a license or legal tool to make their material open (this is discussed further in **1.2**).

1.1.2 - Aid information is protected by different legal rights

Aid information can come in different forms - from official documents, reports and minutes of meetings, to spreadsheets of financial information and databases of geographical locations. These different forms can be protected by different kinds of legal rights.

Copyright is a property right that covers certain types of works, including most creative and artistic works. In many jurisdictions there may also be rights in data. For more on copyright, see **4.1.1**. For more on legal rights in data, see **4.1.2**.

1.1.3 - Examples of legal rights in aid information

Following is a table of examples of different kinds of documents and datasets and the legal rights they may be protected by:

Example	Copyright	Rights in data
A document with text and images	Copyright in text and images.	
A document with text, images and an appendix containing tables of financial information	Copyright in text and images. May be copyright in	

1. See also the 'Many Minds Principle', in http://m.okfn.org/files/talks/xtech_2007/

	selection and arrangement of information in tables.	
A spreadsheet detailing financial transactions, with fields for recipient name, transaction amount and date	May be copyright in selection and arrangement of information in spreadsheet.	May be protected by data rights in some jurisdictions.
A spreadsheet with details of donor organisations, descriptions of these organisation, and geographical information	Textual descriptions of organisations may be protected by copyright. May be copyright in selection and arrangement of information in spreadsheet.	May be protected by data rights in some jurisdictions.
A relational database with details of aid projects, project donors and project beneficiaries, including dates, amounts, descriptions, contact details, geographical locations and notes.	Textual descriptions and notes may be protected by copyright.	May be protected by data rights in some jurisdictions.
A website with text, photos, and video material, along with datasets available for download.	Copyright in text, images and video.	Datasets may be protected by data rights in some jurisdictions.

1.2 - Legal mechanisms for allowing material to be re-used and re-distributed

1.2.1 - By default re-use requires permission

Given these legal rights, any third party who wished to reuse or re-distribute either content or data pertaining to international development could be prevented from doing so or face uncertain legal risk.

By default, organisations who publish aid information may grant (or deny) permission to reuse their material on a case-by-case basis. This case-specific process is expensive in terms of time and resources for both publishers and prospective re-users. Submitting and processing re-use requests may take up person hours that could be used elsewhere.

Also anyone wishing to re-use a collection of material where permission has been sought individually for each item, may themselves have to re-seek permission for each item. For example, say John has sought permission to aggregate, harmonise and re-publish demographic statistics gathered by various NGOs and local authorities. If Sarah wishes to combine and re-publish the same material with material she has collected she will need to ask permission from each rights holder again.

1.2.2 - Using a license or legal tool to facilitate re-use

If those publishing aid information wish to allow others to utilise their material without having to ask permission, they can use a license or legal tool which makes permissions, restrictions and conditions of use clear. Organisations can choose to apply a license to their material that clearly allows others to reuse it without asking for further permission.

1.2.3 - 'Bespoke' and 'generic' licenses and legal tools

These licenses and legal tools may be bespoke, drafted for a specific person, project or organisation, or they may be generic, intended to be re-used in a wide variety of contexts.²

On the one hand someone who publishes content or data may include a statement which says what others are allowed and not allowed to do with the material. For example, they may state that anyone can use the material as long as they cite the original authors, or that anyone can use the material but only for particular purposes (for personal use, for educational purposes, and so on).

The UK's Click Use PSI License is a bespoke license, specifically drafted to facilitate the re-use of core UK government Public Sector Information (PSI)³. It is meant to be used for Crown Copyright material from the UK government, and is not intended to be used by other parties.

On the other hand there are 'generic' licenses and legal tools which are meant to be used by anyone. Creative Commons has developed a set of licenses to help any creator allow others to utilise their works as long as they comply with certain conditions - such as attributing the creator, not using the material for commercial purposes, or sharing any derivative works under the same or a similar license.⁴ Open Data Commons has developed a set of legal tools aimed specifically at data⁵.

2. A similar distinction, between 'custom' and 'standard' licenses, is made by ccLearn in their *What status for open?* report: <http://learn.creativecommons.org/what-status-for-open.html>

3. See <http://www.opsi.gov.uk/click-use/index>

4. See <http://creativecommons.org/license/>

To give an example, Sarah may publish an article she has written under a Creative Commons Attribution license. This allows John to re-use the article in any way he likes as long as he explicitly credits Sarah. Just as important, the CC license allows Fred, after receiving Sarah's article in John's use to also re-use Sarah's article.

These generic licenses are now in widespread use by NGOs, media organisations and government agencies⁶ as well as individual creators.

1.2.4 - Different kinds of sharing: from accessibility to openness

There are a wide variety of different generic licenses available which build on copyright and stipulate different conditions that prospective re-users must comply with.

These range from minimally permissive licenses which, for example, simply allow verbatim re-distribution or copying for personal, noncommercial purposes, to fully 'open' licenses which impose no restriction on use or reuse other than, perhaps, the requirement to attribute or share back.

As well as asserting rights and permitting re-use given certain conditions, creators can alternatively make a statement to the effect that they waive their rights or dedicate material to the public domain - in effect allowing anyone to re-use it with no restrictions whatsoever.⁷

In order to encourage third parties to re-use and build on content and data, we strongly recommend the use of a license or legal tool which complies with the Open Knowledge Definition (<http://www.opendefinition.org>). The Definition sets out principles to define 'openness' in content and data. It requires that - in addition to access - re-distribution and re-use are permitted.

1.2.5 - Making content and data open

As discussed in 1.1, there are different kinds of rights in different materials. Hence different legal mechanisms are required in order to make these materials open.

5. See <http://www.opendatacommons.org/licenses/>

6. For example, see http://wiki.creativecommons.org/Government_use_of_CC_licenses

7. For example the Public Domain Dedication & License (<http://www.opendatacommons.org/licenses/pddl/1.0/>) and CCo (<http://creativecommons.org/publicdomain/zero/1.0/>)

There are licenses designed to be applied to copyrighted content such as the Creative Commons Attribution or Attribution-Sharealike licenses, as well as legal tools meant for data, such as the Public Domain Dedication & Licence (PDDL), the Open Database License (ODbL) or CCo. A comprehensive list of licenses and legal tools⁸ for making data open can be found at the Open Definition site, as well as a detailed guide to open licensing for both types of material.⁹

While publishers can use a bespoke license (as discussed in **1.2.3**), we would strongly recommend that publishers use one of a small number of existing open licenses/legal tools intended for widespread use. These would be stable, actively maintained, and interoperable - ensuring material under one license can be combined with material under another.

Following is a table outlining a basic set of open licenses and legal tools for content and data:

License/legal tool	For content?	For data?	Attribution requirement?¹⁰	Sharealike requirement?¹¹
Creative Commons Attribution	✓	X	✓	X
Creative Commons Attribution Sharealike	✓	X	✓	✓
Public Domain Dedication & License (PDDL)	X	✓	X	X
Open Database License (ODbL)	X	✓	✓	✓
Creative Commons Zero	✓	✓	X	X

8. See <http://opendefinition.org/licenses/>.

9. See <http://opendefinition.org/guide/>.

10. I.e. the requirement that anyone using the material should attribute its creators.

11. I.e. the requirement that anyone using the material should license any derivative works under the same (or similar) license.

1.3 - Other law and policy pertaining to sharing aid information

1.3.1 - Law and policy related to sharing core government information

Governments may be legally or politically obliged to permit re-use and re-distribution of certain core government material. In 2008 the OECD recommended that its members make Public Sector Information (PSI) open by default:

Openness. Maximising the availability of public sector information for use and re-use based upon presumption of openness as the default rule to facilitate access and re-use.¹²

In Europe the EU PSI Directive encourages member states to allow material to be re-used¹³. While implementation of the Directive is ongoing, several governments have responded by making their information assets open. For example, in the UK, core government material is made available under a Click Use License¹⁴ (as mentioned in **1.2.3**) which is compliant with the *Open Knowledge Definition*.

In the United States Federal government material by law cannot be copyrighted, and hence is in the public domain. Many U.S. government agencies holding data relevant to aid - such as the U.S. Agency for International Development - are obliged to release their data into the public domain¹⁵. (Though it should be noted that this is not necessarily the case for U.S. State and Local government.)

Furthermore, several governments, including the US (for non-public domain content) and Australia, have used open licenses such as the Creative Commons Attribution license for their material¹⁶.

12. See OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, especially p. 5ff. <http://www.oecd.org/dataoecd/0/27/40826024.pdf>

13. See http://ec.europa.eu/information_society/policy/psi/actions_ms/implementation/index_en.htm

14. See <http://www.opsi.gov.uk/click-use/index>

15. Such as data in the U.S. Agency for International Development's Economic Analysis and Data Services (EADS): <http://ckan.net/package/read/usaid-eads>

16. See http://wiki.creativecommons.org/Government_use_of_CC_licenses

1.3.2 - Freedom of Information (FOI) laws

While Freedom of Information (FOI) laws permit citizens to request access to official documents and datasets - these may not always give permission for the information requested to be republished or re-used. This may limit the extent to which information can be recombined and represented using digital technologies. As discussed in **1.2.4**, we strongly suggest that information related to development be made open, not just accessible.

1.3.3 - Privacy laws

Depending on the type of aid information, there may be privacy issues. In this paper, we are primarily referring to non-personal information.

- **Recommendation 1 - Aid information should be legally open.** The standard should require a core set of standard licenses for publishing aid information under. It should require that either:
 - (i) information is published under one of a small number of recommended options:
 - Licenses for content: Creative Commons Attribution or Attribution Sharealike license
 - Legal tools for data: PDDL, ODbL or CCo
 - or that (ii) information is published using a license/legal tool that is compliant with a standard such as the *Open Knowledge Definition*.

2. Making aid information technologically open

2.1 - File formats

Aid information can be published in a wide variety of different file formats - from word processor document formats to database files formats. The choice of file format can affect how easy it is to re-use the information encoded in it.

2.1.1 - Proprietary, non-proprietary and open file formats

On the one hand some file formats are non-proprietary and open, which means they can be used or implemented by anyone with little or no restriction. Prominent examples include HTML/XHTML, OpenDocument, PDF, TXT, XML¹⁷.

On the other hand some file formats are proprietary, which means that there may be restrictions on how the format may be used, and certain software packages may be required to read the files. Prominent examples include MPEG Audio Layer 3 (MP3), Windows Media Video (WMV), Microsoft Word (DOC/DOCX) and Microsoft Excel (XLS/XLSX).¹⁸

2.1.2 - Machine readability

While some file formats present data in a way which is 'machine-readable', some file formats are primarily meant to be read by people.

For example, a table of financial information might be published in XML, Comma Separated Value (CSV) or Microsoft Excel (XLS) formats, which can be easily graphed, analysed, aggregated with other data or converted into other formats - as the rows and columns can be read by the computer. Alternately the same table could be published in a PDF or Microsoft Word Document. In this case the material in the tables would have to be extracted - either by hand or by using a computer program (commonly called 'screen scraping'). Also data may be broken up into multiple tables and scattered throughout a document with explanatory notes - in which case the underlying data would have to be extracted and pieced back together.

17. See list at <http://www.openformats.org/en9>

18. Both Word and Excel file formats are included on the Microsoft Open Specification Promise <http://www.microsoft.com/interop/osp/> published in September 2006 - which is a promise not to assert legal rights over certain formats.

It is worth noting that whether or not a file format is open is a separate issue from whether or not it is machine readable. For example, PDF files are open - but the file format is mainly orientated towards printing and layout, not for their contents to be extracted, revised and/or re-used after publication. Excel files are proprietary but they are machine processable and are much more useful than PDF files when it comes to analysing, visualising or linking together their contents.

Following is a list of common file formats for text and data along with details on whether or not they are machine readable, whether the specification is available and whether or not they are open¹⁹:

File format	Machine readable?	Specification available?	Open?
Plain Text (.txt)	✓	✓	✓
Comma Separated Value (.csv/.txt)	✓	✓ ²⁰	✓
Hyper Text Markup Language (.html/.htm)	✓	✓	✓
Extensible Markup Language (.xml)	✓	✓	✓
Resource Description Framework (.rdf)	✓	✓	✓
Open Document Format (.odt, .ods, etc)	✓	✓	✓
Portable Document Format (.pdf)	X	✓	✓
Microsoft Word (.doc/.docx)	X	✓	X
Microsoft Excel (.xls/.xlsx)	✓	✓	X

19. This is based on a registry of formats from the Information Accessibility Initiative: <http://wiki.okfn.org/p/iai/registry>

20. Though there is not a official standard specification for CSV, many informal specification documents exist.

2.1.3 - Formats for the semantic web

There are a variety of formats designed to make it possible for computers to analyse the contents of a file. For example, Resource Description Framework (RDF) allows resources to be described in such a way that computers can sort and query their contents.

So while a encyclopedia might contain an ordinary sentence such as "Paris is the capital of France" which means nothing to a computer, an equivalent statement in RDF could express that 'Paris' is the name of a capital city, 'France' is the name of a country, and that the first 'is a capital of' the second. Hence, RDF allows the creation of structured relationships between entities that computers can parse and query - rather than unstructured text that the computer can do much less with. As a concrete example, projects such as DBpedia attempt to extract structured information from Wikipedia to allow users to make sophisticated queries such as:

[...] soccer players with number 11 (on their jersey), who play in a club whose stadium has a capacity of more than 40000 people and were born in a country with more than 10 million inhabitants.²¹

2.1.4 - Which formats?

It may be difficult to know which formats will be most useful in the long term - and hence it may be undesirable to prescribe a single format which may be popular or widespread at one point, but superseded in the future. This may incur unnecessary costs to organisations publishing the material, and could require expertise that is unavailable.

When it comes to re-using information, it is crucial that the data is machine readable and that there are no technical obstacles to re-using it. If it is under an open license, then others can republish the same material in different formats. If the file format specification is publicly available, then there is less risk that prospective re-users will be required to use a particular piece of software, or that, in the worst case scenario, the format will become obsolete and unreadable without software that is no longer supported.

While publishing information in formats such as RDF and JSON would be desirable, not all organisations will have the expertise to do this. Hence we would not recommend that the information publishing standard should require it.

21. Sören Auer, Jens Lehmann: What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. In Franconi et al. (eds), Proceedings of 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, 2007, LNCS 4519, pp. 503–517, ISBN 978-3-540-72666-1, Springer, 2007. See: <http://www.informatik.uni-leipzig.de/~auer/publication/ExtractingSemantics.pdf> p.11

We suggest that the standard should require that aid information is published in one of a small number of basic formats for text and data such as TXT and CSV, and recommend (optionally) that it is also published in other more recent formats such as RDF, JSON and so on. Crucially it should require that all formats are machine readable.

2.2 - Data schemas for aid information

2.2.1 - Standard and non-standard schemas

A data schema structures and describes information of a certain kind. Schemas may be non-standard, specific to a particular application or purpose, or they may adhere to existing standards.

For example, Jessica may decide to create a catalogue of her books. She might devise her own way of doing this using database software - and make her own decisions about how to catalogue different types of publications, how to record different pieces of information, what information to include and so on. Alternately, she might create her catalogue in adherence to the Machine Readable Cataloging (MARC) standards - which specify how she should catalogue her books, and give a large range of standard fields which she can use as well as details about how the information should be formatted²². In the latter case, she will be able to import bibliographic records adhering to the MARC standards, and share her records with others that use MARC. In the former case, the catalogue will most likely not be interoperable with other catalogues - unless perhaps work is done to translate the records in her catalogue to a standard format.

Adhering to a standard data schema is important to ensure that datasets can be compared and combined together. Without adhering to a standard, significant work may have to be done to harmonise data from different sources.

2.2.2 - Existing schemas and standards

There are many data schemas in different domains that may be of interest to those publishing aid information.

In the late 1990s, the International Development Markup Language (IDML) was created to assist with the exchange of international development related information.²³ The Development Gateway became interested in the project in 2001, and are currently developing it.²⁴

22. See: <http://www.loc.gov/marc/>

23. See <http://xml.coverpages.org/dml.html>

24. See <http://www.dgfoundation.org/fileadmin/templates/pdfs/accraidml.pdf>

In relation to statistical data, there is the Statistical Data and Metadata eXchange (SDMX) standard, which is supported by many national and international organisations, including the United Nations, the OECD, Eurostat, World Bank, the IMF, and many others.

In relation to financial and business information, schemas include the Financial Information eXchange (FIX), the Financial products Markup Language (FpML) and the eXtensible Business Reporting Language (XBRL).

In addition to domain specific schemas, there are also cross domain initiatives, such as the Dublin Core standard²⁵, which develops metadata standards to facilitate sharing information.

2.2.3 - Which schemas?

We propose that further work should be done towards a data schema for the exchange of information related to international development. Such work should take into account existing developments, and collaborate with relevant parties to ensure interoperability with existing schemas and compliance with existing standards.

2.3 - Accessing aid information

2.3.1 - Access via an API

There are many different ways to publish aid information online. It may be published on web pages, exposed via a query interface on a website, or via an Application Programming Interface (API). Making information available via an API can ensure that it is up to date - or even available in real time.

We suggest that aid information should be published via an API where possible.

2.3.2 - Access to raw data

We strongly advise that raw data be made available in addition to being accessible via an API. This allows others to use the data beyond read-only access. For example, it allows batch conversion into a new format, linking and integration with other resources, archival in multiple places, and so on²⁶. While the latest version of the data may be made available via an API, raw data should be made available in bulk at regular intervals.

25. See <http://dublincore.org/>

26. See <http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/>

For example, the Eurostat statistical service has a bulk download facility offering over 4000 data files. It is updated twice a day, offers data in Tab Separate Value (TSV) format, and includes documentation about the download facility as well as about the data files.²⁷

Another example is the District of Columbia OCTO's Data Catalogue, which allows data to be downloaded in CSV and XLS format, in addition to live feeds of the data.²⁸

We suggest that the standard should require aid information to be made available in raw form on a regular basis.

- **Recommendation 2 - Aid information should be technically open.**
The standard should require that raw data is made available in bulk (not just via an API or web interface) with any relevant schema information and either:
 - (i) in one of a small number of recommended formats:
 - Text: HTML, ODF, TXT, XML
 - Data: CSV, XML, RDF/XML
 - or (ii) in a format:
 - (a) which is machine readable and
 - (b) for which the specification is publicly and freely available and usable

27. See: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/bulk_download and <http://ckan.net/package/read/eurostat>

28. See: <http://data.octo.dc.gov/>

3. Making aid information easy to find

In order for information to be re-used, prospective re-users have to know it is there and to be able to find it easily.

3.1 - Publishing on the web

At the very least this means it should be published on the web, linked to by relevant sites and indexed by search engines. This is already the case for a great deal of aid information, and yet it is still far from straightforward to locate. Effort is required to trawl through, analyse and describe documents and datasets.

This process could be made easier if aid information were marked with a common piece of code or text that could be searched for. For example, aid information publishers might use a web button or distinctive link or piece of text. Additionally, the resource could be appropriately described using RDF or RDFa. However this still requires using a search engine to trawl for relevant datasets, and the ways in which data can be searched and explored may be limited.

3.2 - A registry of aid information

Hence it would be highly desirable to also have a basic registry of aid information assets to make it as easy as possible for people to discover what is available.

3.2.1 - Metadata for datasets

Such a registry would require some metadata describing each resource. Aid information is diverse and can range from geospatial information to budgetary data. Developing a registry would require specifying the lowest common denominator of metadata which will be applicable to any given item.

As a bare minimum we would suggest:

- Short name (as a unique identifier and for URL)
- Title/name of resource
- URL
- Download URL
- License/legal status
- Notes

Other metadata could be added as required or optional fields - preferably adhering to some common metadata standard such as Dublin Core²⁹. For example, for datasets pertaining to a particular country, a country field could be added. For geospatial data metadata may include a bounding box of the region the data pertains to. For financial records pertaining to a certain time period, this information may be specified. The OECD has published a report and details of metadata scheme for publishing datasets that could be informative³⁰.

3.2.2 - Ontologies and tags: user driven resource description

Tags are a useful way of describing large collections of diverse knowledge resources. Many popular web services such as Flickr, Delicious and Twitter have found tagging to be useful to help their users explore and navigate the large amounts of content they produce.

In the past knowledge resources have been described according to some 'ontology' - a fixed, predetermined range of fields used to describe the resource, including names, dates, descriptions, relationships between items being described and so on. By contrast 'social' web services often allow anyone to add any tag, or keyword, to describe a given resource. These keywords do not need to be in accordance with some pre-defined scheme.

For example, on the photo website Flickr, users can add appropriate tags to describe photographs, which can be anything from the names of people, places, events or objects depicted, observations about the subject matter, and so on. Whereas a photo catalogue with a fixed set of fields may only contain information in accordance with these fields (perhaps including the name, date and photographer of a given picture), social tagging arguably allows a much richer spectrum of description. These tags facilitate with the discovery of the items they describe - for example by allowing people to search for all the photographs tagged with 'red sky', taken in Reykjavik, or depicting a certain brand of car.

Many suggest that tags are more flexible and dynamic than a pre-specified ontology - as they can be added, updated or revised 'on the fly'. We think there may be significant benefits to allowing users to tag, or otherwise describe, information resources related to international development.

29. See <http://dublincore.org/>

30. See "We Need Publishing Standards for Datasets and Data Tables"
[http://www.oecd.org/document/25/
0,3343,en_21571361_33915056_42600857_1_1_1_1,00.html](http://www.oecd.org/document/25/0,3343,en_21571361_33915056_42600857_1_1_1_1,00.html)

3.2.3 - Linking, mirroring and storage

The registry would contain links to documents and datasets containing aid information. This material would be stored externally - either on the websites of the organisations publishing the information, or by third parties who specialise in storing such material (such as the Internet Archive or Talis' Connected Commons³¹). As the material would be open, others would be free to deposit it in external archives, mirror (or host copies of) it elsewhere, or to aggregate it with other material.

3.2.4 - Making the registry open

We recommend that the information in the registry should be published under an open license so that others can build on it and integrate it with other resources³². To give a recent example of where this has been useful, the Sunlight Foundation recently announced that it will build on the recently launched data.gov³³.

3.3 - A decentralised model for aid information

We suggest that the standard adopts a decentralised model for the distribution and storage of aid information - rather than one in which content and data is deposited in some central repository.

There are several advantages associated with a decentralised approach:

- The costs associated with maintaining and updating a central repository could potentially be high. Using a registry which users can edit can reduce ongoing costs.
- Multiple versions of the information can be stored in multiple locations, such as organisational websites, third party websites (individuals, other organisations, ...), third party storage facilities (e.g. the Internet Archive, Amazon Web Services, Talis Connected Commons, ...) and so on.
- Data publishers may wish to (continue to) primarily publish and syndicate their data on their own website, rather than in a third party repository. E.g. government departments and international organisations may have their own established systems and procedures for publishing data.
- This model allows users to build on datasets, and contribute improved versions back to the 'ecosystem' (such as converting data to particular formats such as RDF, aggregating datasets and so on).

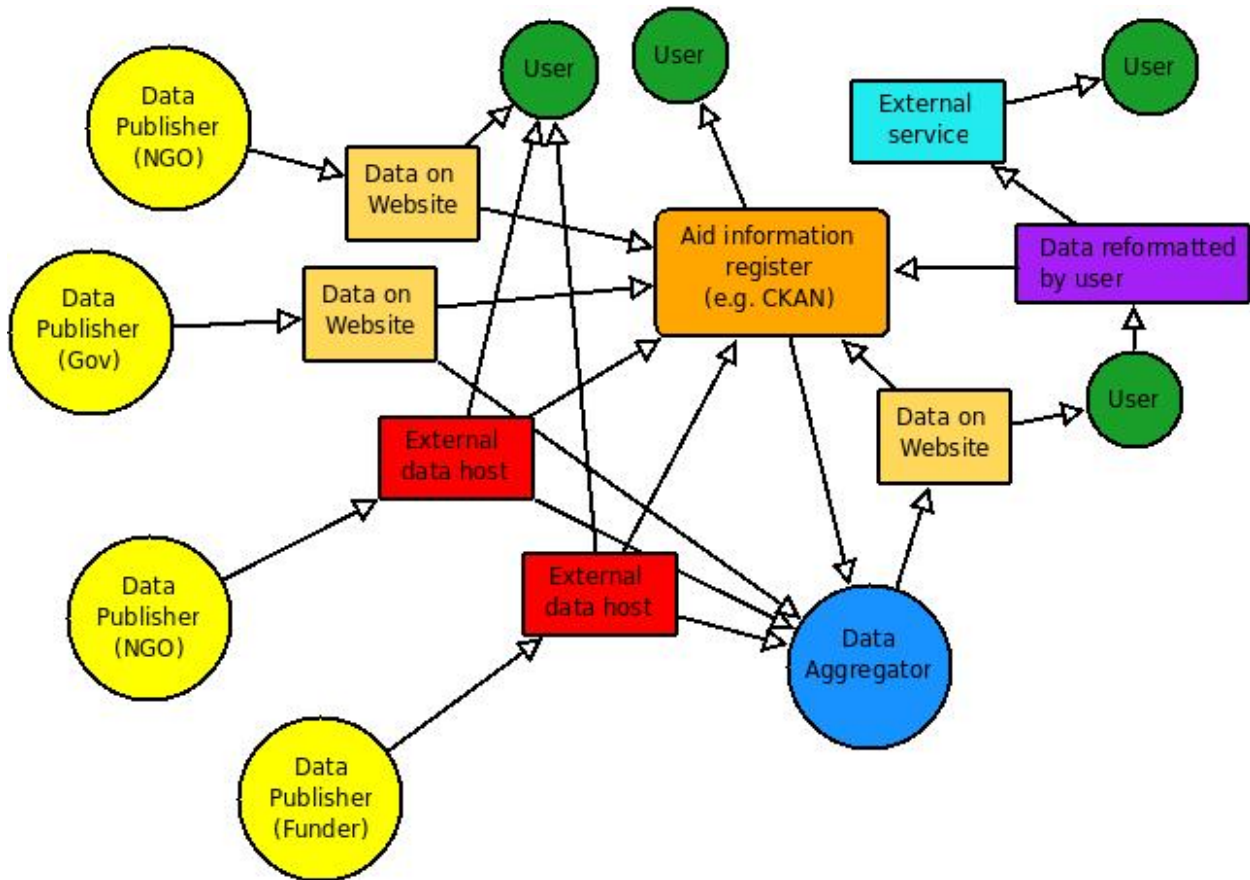
31. See <http://www.archive.org> and <http://blogs.talis.com/n2/cc>

32. See also <http://blog.okfn.org/2009/07/23/what-features-should-be-included-in-a-catalogue-of-open-government-data/>

33. See <http://blog.sunlightfoundation.com/2009/07/15/the-national-data-catalog/>

- It is suited to scenarios in which there are many actors, with different needs from the data.
- It is suited to the incremental, collaborative development of data.³⁴

Below is a diagram representing how information might move around between data publishers, users, websites and storage services.



- **Recommendation 3 - Aid information should be easily findable.** The standard should require that aid organisations add their knowledge assets to a registry with a minimum of basic metadata for each item. The data in the registry should be made open.

34. See: <http://blog.okfn.org/2006/05/09/the-four-principles-of-open-knowledge-development/> and <http://blog.okfn.org/2007/04/30/what-do-we-mean-by-componentization-for-knowledge/>

4. Appendix

4.1 - Legal rights in content and data

4.1.1 - Copyright

Copyright is a property right that covers certain types of works, including most creative and artistic works such as paintings, sculpture, literature, films, television, and music. Worldwide, the vast majority of countries have copyright laws that meet certain minimum standards set out in a number of international treaties.

Copyright grants a monopoly to the rights holder over doing certain acts with the work, including to:

- Reproduce the work (make copies);
- Distribute the work to the public;
- Rent or lend the work to the public;
- Publicly perform the work;
- Broadcast the work or include it in a cable television service; and
- Adapt the work or to do any of the above with an adaptation of the work.

The rights owner of a copyrighted work can thus prohibit others from doing any of the above acts, unless an exception or limitation to copyright applies. Use of the work in ways not covered by fair dealing or other exception requires permission from the copyright holder. Permission to use a copyrighted work usually comes in the form of a licence, which is a legal document outlining what can and can't be done with the work.

As a general rule, copyright on texts, images, audio, and video automatically vests (without the need for registration) with their creator or, in the case of employment, with the employer. In the context of aid information – copyrighted material may include written reports, illustrative graphs, audio recordings of speeches or broadcasts, contents of emails or text on a website.

4.1.2 - Rights in data

Factual information by itself is not covered by copyright, but several different legal rights can cover databases of factual information including copyright. For example in the EU, the Database Directive (Directive 96/9/EC) grants a specific separate "database right" for database creators, as well as copyright in the way in which data or other material is collected, selected and arranged in the database³⁵. Contract and other rights such as technical protection measures (often referred to as DRM) can also cover

databases. As a result, databases of aid information – from financial records to geographic locations – will also be covered by legal rights.

In both copyright and database rights certain exceptions apply, such as fair use or fair dealing. However, the default of automatic protection of both creative works (text, images, audio) and of databases means that – unless clearly covered by an exception – subsequent users of those works and databases will need to seek permission before use. This is sometimes referred to as an "all rights reserved" model: the cloud of automatic legal restrictions and unknown legal risk for re-use lead to a "permission culture" where almost all use and re-use is predicated with seeking formal permission from the rightsholder.

4.2 - Existing examples of aid information

Following are several examples of existing online aid information, along with details of:

- **Legal openness.** A review of any licenses, legal statements and terms of use. Details of whether they comply with the *Open Knowledge Definition*. (See 1.2)
- **Open format.** Is the data in an open format? (See 2.1)
- **Machine readable format.** Is the data in a format which is machine readable? (See 2.1)
- **Bulk downloadability.** Is the raw data is offered for download in bulk (in addition to access via an API, if applicable)? (See 2.3)

These are summarised in the following table:

Projects	Legally open?	Open format?	Machine readable format?	Bulk download available?
World Bank Data (4.2.1)	X	X	✓	X
Asian Development Bank (4.2.2)	X	✓	X	X
USAID (4.2.3)	✓	✓	✓	✓
World Food Programme - Food Aid	?	✓	✓	X

35. For more information on rights in data in different jurisdictions, see the 'Guide to Open Data Licensing' <http://opendefinition.org/guide/data/>

Information System (4.2.4)				
European Commission Humanitarian Aid (ECHO) - Funding (4.2.5)	✓	✓	X	✓
Organisation for Economic Cooperation and Development (OECD) Development Assistance Committee (DAC)	X	✓	✓	X

For more examples, refer to the list of packages tagged 'development' on CKAN:

- <http://ckan.net/tag/read/development>

4.2.1 - World Bank Data & Statistics

URLs

- <http://www.worldbank.org/data>
- <http://ckan.net/package/read/world-bank>

Legally open?

No. Commercial re-use not allowed. Terms of use say:

The World Bank is pleased to allow Users to visit the Site and download and copy the information, documents and materials (collectively, "Materials") from the Site for User's personal, non-commercial use, without any right to resell, redistribute or create derivative works therefrom, subject to the terms and conditions outlined below, and also subject to more specific restrictions that may apply to specific material within this Site.

Open format?

No. The specification for XLS is publicly accessible, but it is not open.

Machine readable?

Yes. XLS is machine readable.

Bulk download available?

No bulk download facility.

4.2.2 - Asian Development Bank

URLs

- <https://sdfs.adb.org/sdfs>
- <http://ckan.net/package/read/adb-sdfs>

Legally open?

No. Reproduction prohibited. Page footer says:

Copyright © 2007 Asian Development Bank. All rights reserved. Reproduction in whole or in part without permission is prohibited.

Open format?

Yes. Material is available in PDF.

Machine readable?

No. PDF is not machine readable.

Bulk download available?

No bulk download facility found.

4.2.3 - U.S. Agency for International Development - Economic Analysis and Data Services (EADS)

URLs

- <http://www.usaid.gov/km/eads.html>
- <http://ckan.net/package/read/usaid-eads>

Legally open?

Yes. The FAQ states:

Unless a copyright is indicated, information on this Web site is in the public domain and may be reproduced, published or otherwise used without USAID's permission. We request that USAID be cited as the source of the information and that any photo credits or bylines be similarly credited to the photographer or author or USAID, as appropriate.

If a copyright is indicated on a photo, graphic, or any other material, permission to copy these materials must be obtained from the original source.

Open format?

Yes. Available in HTML, which is an open format.

Machine readable?

Yes. Available in HTML and XLS, both of which are machine readable.

Bulk download available?

There is not a single bulk download facility, but the query interfaces have an option to download material for all countries.

4.2.4 - World Food Programme - Food Aid Information System

URLs

- <http://www.wfp.org/fais/>
- <http://ckan.net/package/read/wfp-fais>

Legally open?

No explicit information found.

Open format?

Yes. Exports to Comma Separated Value (CSV) format.

Machine readable?

Yes. Exports to Comma Separated Value (CSV) format.

Bulk download available?

No facility found.

4.2.5 - European Commission Humanitarian Aid (ECHO) - Funding

URLs

- http://ec.europa.eu/echo/funding/finances_en.htm
- <http://ckan.net/package/read/eu-echo-funding>

Legally open?

Yes. Free to be re-used as long as attribution is given. The copyright notice on the European Commission website states:

Reproduction is authorised, provided the source is acknowledged, save where otherwise stated. Where prior permission must be obtained for the reproduction or use of textual and multimedia information (sound, images, software, etc.), such permission shall cancel the above-mentioned general permission and shall clearly indicate any restrictions on use.

Open format?

Yes. Available in PDF, which is an open format.

Machine readable?

No. PDF is not machine readable.

Bulk download available?

Yes. PDFs all linked to on a single page.

4.2.6 Organisation for Economic Cooperation and Development (OECD) Development Assistance Committee (DAC)

URLs

- <http://www.oecd.org/dac/stats/data>
- <http://ckan.net/package/oecd-ids>

Legally open?

No. Terms and conditions from main OECD site mention non-commercial restrictions.

You can copy, download or print content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. You should cite the Title of the material, © OECD, publication year (if available) and page number or URL (uniform resource locator) as applicable. Please consult the Rights and Permissions section at www.oecd.org/rights for all requests for public or commercial use and translation rights.

Open format?

Yes. Available in CSV, which is an open format.

Machine readable?

Yes. Available in XLS and CSV - both of which are machine readable.

Bulk download available?

No. Individual datasets can be downloaded in XLS or in CSV format - but there is no single directory or zipped file containing all datasets in downloadable form. Instead user has to navigate to each data theme individually and click to export each dataset individually via the web interface.

4.3 - Existing Work

4.3.1 - Legal and licensing standards

The Open Knowledge Definition (OKD) sets out principles to define the 'open' in open knowledge. The term knowledge is used broadly and it includes all forms of data, content such as music, films or books as well any other type of information.

In the simplest form the definition can be summed up in the statement that "A piece of knowledge is open if you are free to use, reuse, and redistribute it".

For full details see:

- <http://opendefinition.org>

4.3.2 - Format standards

There are a plethora of definitions of what an 'open standard' is - from the legal definitions of international standards organisations and national governments to definitions from companies, NGOs and influential technologists.

For further information see:

- <http://www.openformats.org/main>
- http://en.wikipedia.org/wiki/Open_standard
- http://en.wikipedia.org/wiki/Open_format

A list of file formats, including information on which are machine readable, is available at:

- <http://wiki.okfn.org/projects/iai/registry>

For web publishing standards, see the World Wide Web Consortium:

- <http://www.w3.org/>

4.3.3 - Data and metadata standards

4.3.3.1 - Resource description

- CKAN metadata. Minimum of required metadata for open datasets. See:
- CKAN: <http://ckan.net/>

- Example package: <http://www.ckan.net/package/read/at-budget>
- Dublin Core <http://dublincore.org/>
- Suggestion for OECD dataset metadata standard http://www.oecd.org/document/25/0,3343,en_21571361_33915056_42600857_1_1_1_1,00.html

4.3.3.2 - Data models

There are also a range of standards for publishing different kinds of data, including:

- Statistical Data and Metadata Exchange (SDSX): <http://sdmx.org/>
- International Development Markup Language (DML): <http://xml.coverpages.org/dml.html>
- Financial Information eXchange (FIX): <http://www.fixprotocol.org/>
- the Financial products Markup Language (FpML): <http://www.fpml.org/>
- the eXtensible Business Reporting Language (XBRL): <http://www.xbrl.org/Home/>

4.3.4 - Other relevant work

- Josh Tauberer, "Open Government Data", <http://razor.occams.info/pubdocs/opedataciviccapital.html>
- "Improving Access to Government through Better Use of the Web", W3C Interest Group Note, 12 May 2009 <http://www.w3.org/TR/2009/NOTE-egov-improving-20090512/>
- Power of Information Review
- Tim Berners-Lee, "Putting Government Data online", <http://www.w3.org/DesignIssues/GovData.html>