This is the PDF version of the original document at
http://philarcher.org/diary/2012/linkedopenclassification/

Linked Open Classification

This brief document outlines a simple proposal: that classification data concerning films, games etc. as published by bodies like NICAM, PEGI and the BBFC should be made available, for free, through a common platform and under a common licence that allows re-use.

I make this proposal as an individual. It has no standing within W3C, however, depending on future developments, it is possible that W3C would take an interest. This is explored below.

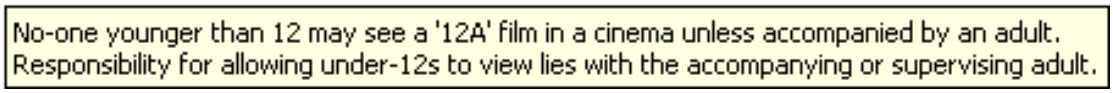
The End Goal

A key strength of the Web is that the barrier to entry is minimal. Everyone from large scale enterprises to individual amateurs can publish content online. Where published material refers to content that has been professionally classified, those classifications should be available directly.

An end user, whether using a normal Web browser or a proprietary App, should be able to access the classification data and be able to verify its authenticity.

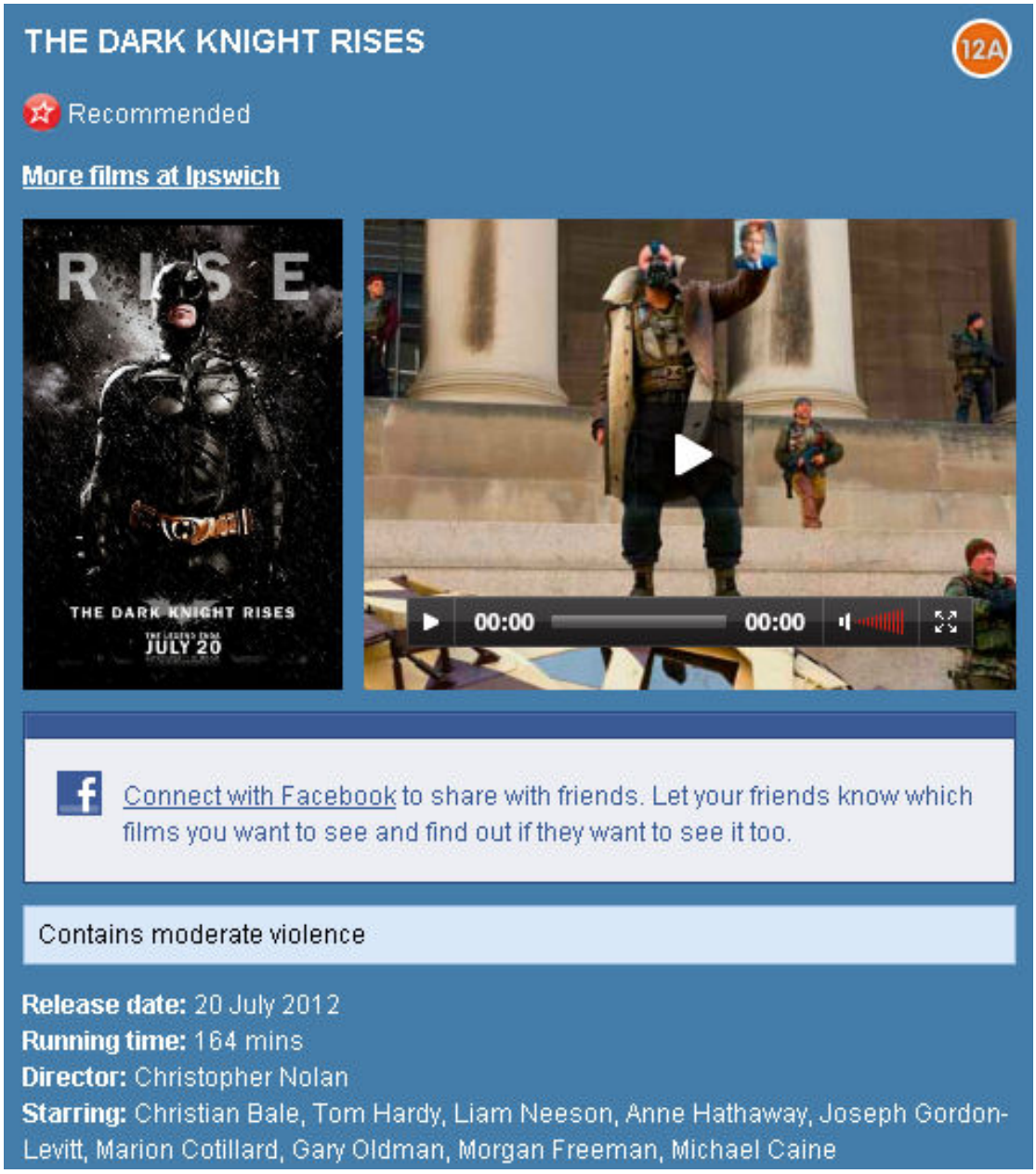
The Dark Knight Rises, as advertised on cineworld.co.uk

As the screenshot above from the Web site of a major UK cinema chain shows, the classification is clearly displayed for human users. If you put your mouse over the classification symbol, you see a bit of explanatory text:



The text displayed when you mouse over the 12A icon

That text is the generic text describing what the rating means and doesn't give any specific information about the film itself. If we go a little deeper into this particular Web site we find a bit more classification information:



The page specifically about this film on cineworld.co.uk. Notice the 'Contains moderate violence' advice

Compare that consumer advice with what's available on the BBFC's own site:

THE DARK KNIGHT RISES

Release date:	20/07/2012
Running time:	164m 31s
Consumer Advice:	Contains moderate fantasy violence

Extended Classification Information is available for this work. ECI may contain plot details and/or spoilers. Learn more » about ECI or Show details »

This work was passed with no cuts made.

Work Information

Director:	Christopher Nolan
Cast includes:	Christian Bale, Joseph Gordon-Levitt, Gary Oldman, Tom Hardy, Liam Neeson, Anne Hathaway, Marion Cotillard, Aidan Gillen, Morgan Freeman, Michael Caine, Matthew Modine, Juno Temple, Tom Conti
Genre(s):	ACTION
Main Language:	English
Distributor:	Warner Brothers Entertainment UK Ltd
Classified:	18/07/2012
Measured footage:	236908 (feet + frames)
BBFC Reference:	CFF281347
Registration number:	CFJ26541

The information available about this film on the BBFC Web site

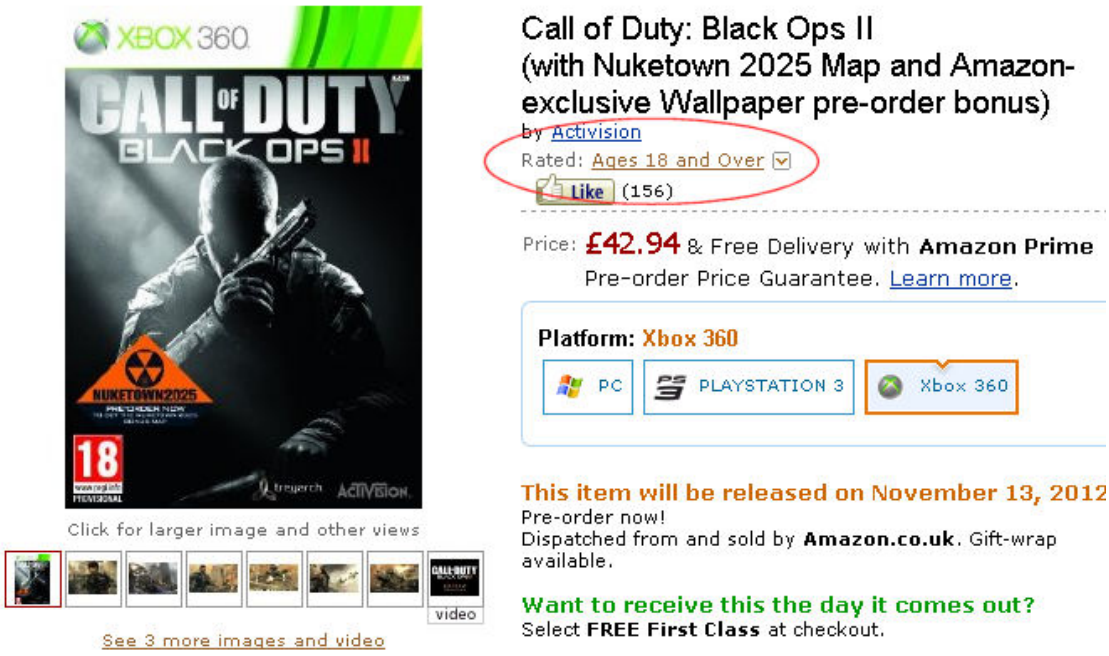
The first thing we might notice is that there is more data available at the BBFC than on the Cineworld Web site, but there's something more important for our discussion here: the consumer advice is different.

There's nothing intrinsically wrong with this. Whether Cineworld chooses to include any consumer advice is entirely a matter for them and there's no rule that says that if they do then it must match what the BBFC says. But a user might reasonably assume that the consumer advice came from the BBFC.

Furthermore, the extended consumer advice available about this film is not included anywhere. To find that, you a) need to know it exists; and b) have to know where to find it on bbfc.org.

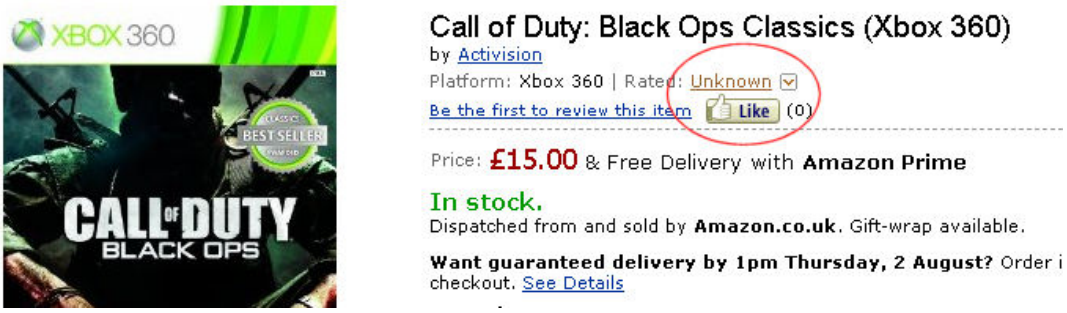
A very similar situation obtains when looking at games. Amazon, for example, shows the rating and then provides just the basic info[1] about what each age classification means, as exemplified by this soon-to-be-released major game.

[1] http://www.amazon.co.uk/gp/help/customer/display.html/ref=dp_vg_help?ie=UTF8&nodeId=502556



The new Call of Duty game as advertised on Amazon.co.uk (screenshot taken 1st August 2012)

Again, for a human user, the PEGI rating information is clear and easily accessed although it is very generic and doesn't inspire confidence that the rating is definitely about that particular game. The situation for the (older, long released) Call of Duty Black Ops is surprising:



Call of Duty Black Ops as sold on Amazon.co.uk with "unknown" classification. (screenshot taken 1st August 2012)

Amazon doesn't appear to know the rating of this game although, of course, it does have a PEGI rating which can be seen at pegi.info.

The PEGI rating for Call of Duty Black Ops

A lot of ratings and ratings-related data is available through the classification boards' websites. You can type in a film or game title, and perhaps other information, to find out what the rating is. Good. But people are lazy and like seeing relevant information where they are rather than having to go to a specific Web site to find it. Even where the classification data is included it is painfully easy to find errors and omissions. The evidence suggests that retailers and review sites don't regard the provision of accurate classification data as a priority. It is something concerned parents will actively look for and the existing search systems operated by the classification boards meet this need. However, if the general consumer is to receive classification information, more effort needs to be made to put it in front of them wherever they are at the time.

This proposal asserts that it would be better if:

- film & game classifications were available through related websites, services and applications;
- the consumer advice and other ancillary information were available at the same time so parents could see the classification in context;
- consumers could know that classification information was genuine, even when seen on third party Web sites;
- developers could include accurate classification data with almost zero effort;
- publishers were able to state that the classification data was 'official' (in the way that PEGI Online and BBFC Online facilitates);
- classifications were accessible from countries other than the the user was in;
- classification information were available as part of the wider datasphere;
- developers were able to use classification data in innovative ways that reached a wider audience.

Crucially, these aims must be met whilst the rating information *remains in the control of the relevant classification agency*.

All of the above can be achieved by:

1. providing a service through which data can be accessed in real time;
2. making classification data available directly as machine readable data under a permissive licence (in addition to existing human-accessible portals).

The Service API

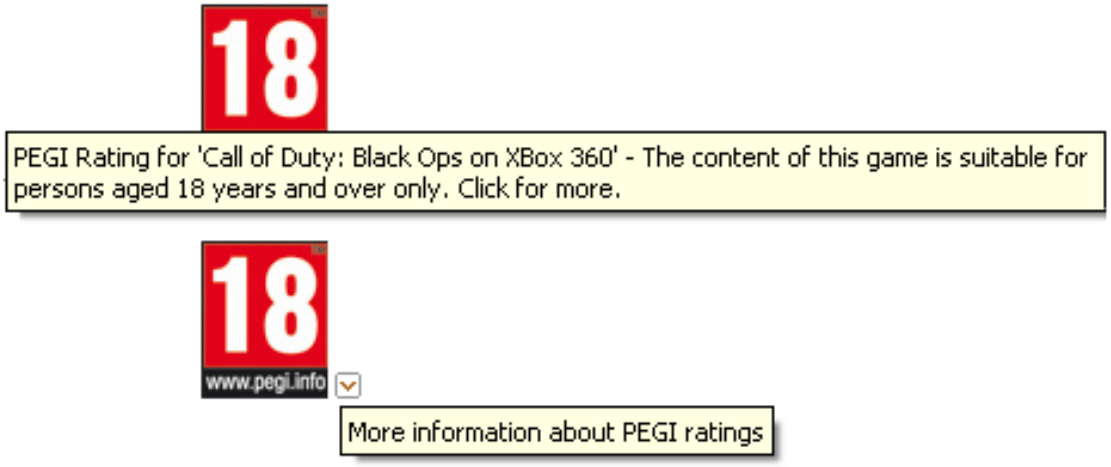
From a developer's point of view, the ideal is to be able to include a minimal amount of code, perhaps like this:

```
includeRating('board:PEGI', 'title:Call of Duty: Black Ops',
'system:Xbox');
```

In other words, write a single line of code that makes a simple request to the service. The response would be the relevant rating information that would be automatically included in a page, complete with access to additional data. Something like this perhaps:

The product page as it could be on Amazon and all retailers, fan/review sites etc.

Placing the mouse over the rating symbol shows that the 18 rating is specific to this particular game. Likewise, placing the mouse over the little drop down symbol shows that more information is available about the PEGI system. (If you're reading this online, you can see this working, the images below are for offline readers.)



The text displayed when the mouse is placed over the rating logo and the drop down

Clicking the rating symbol will bring up the detailed rating in a popup of some kind. For this simple demo, the domain name shown is philarcher.org but in a real implementation, the rating would come from the classification board (in this case pegi.info) and should be served securely over https.

The popup showing the detailed rating information directly from PEGI

In order to make it possible for developers to include this functionality, the classification board needs to make available a very small, almost trivial, code library that can be included in the page and then the developer would add in the kind of product-specific code shown above. This is an approach that developers well understand - things like Twitter and Facebook 'Like' buttons are all included in this way and the general approach is now the subject of standardisation work at W3C known as Web Intents[2]. The 'intent' here would be to find out more about the classification of a film or game.

The less work the developer has to do to create smart functionality on the page — functionality that makes them look good — the happier s/he will be.

Incidentally, the example above is very simplistic and one that doesn't work well on mobile where there is no mouse and popups are a real pain. The code library would need to implement a better solution than shown here but the principle remains the same.

The Data

An API, or service like the one described above would make it easy for regular developers to include classification data directly in their Web pages. Releasing the data as data will reach a different set of developers, those that want to build data-driven applications.

The word 'application' means different things to different people. For many people, Apps are things you download from the Apple or Android App store. For our purposes, an application is any means of visualising and/or interacting with data, usually on a Web site (whether accessed on a desktop or mobile device). This definition includes proprietary apps but covers a lot more besides.

The Open Data Movement

There is a substantial amount of work going in to publishing data. The driver for this has been what one might call the open data movement which is focused on persuading governments to publish their data. This persuasion has worked specticularly well and open data is now a feature of public policy in many countries (see, for example, the UK Government's white paper Unleashing the Potential[3]). Sites like data.gov[4], data.gov.uk[5] and overheid.nl[6] are being replicated rapidly and

---

[2] http://www.w3.org/TR/web-intents/

[3] http://www.cabinetoffice.gov.uk/resource-library/open-data-white-paper-unleashing-potential

[4] http://www.data.gov

[5] http://www.data.gov.uk

not just at national level. Cities such as Chicago[7] are also publishing their data. Many of these portals are instances of a particular platform, CKAN, which is a product of the Open Knowledge Foundation[8] that publishes a list of portals[9] that use its software. There are also a number of commerical platforms designed to allow data holders to monetise their data directly, such as Data Market[10].

What do people do with this data? Well, the short if unhelfpul answer is "all sorts of things." A lot of applications essentially provide a visualisation of a single data set such as Where Did My Tax Go?[11] which is built on the Public Expenditure Statistical Analyses data set[12]. Many applications visualise data by putting them on a map, for example, airText[13] which uses the London Atmospheric Emissions Inventory[14].

The really eye catching applications take data from multiple sources. For example, Pitchup.com[15] uses many data sources to provide an ever-richer Web site about camp sites: their location, nearby attractions, bathing water quality and more. A very different example that uses the same approach is Reegle[16] - a comprehensive source of information about renewable energy.

Open Data and Entertainment

The open data movement sounds all well and good but aren't we supposed to be talking about film and game classification?

There's a problem: there isn't much data to go on.

The image below shows a small section of a diagram that is used pretty well whenever anyone talks about open data. It's from the Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch[17] that shows the manor datasets that have been published as linked data with the size of circle indicating the size of the data set. The colours indicate general themes and the light blue area is of most interest to us here. With the notable exceptions of the BBC programmes data[18] and the Linked Movie Database[19], the data is very much music industry focused. Film, TV and games are all-but absent and the linked MDB, a response to the closed & commercial nature of IMDB, suffers from lack on maintenance. This despite the community demand[20] for just such a databse.

What's missing is an authoritative, industry-lead data set.

---

6 http://overheid.nl/

7 https://data.cityofchicago.org/

8 http://okfn.org/

9 http://ckan.org/instances/

10 http://datamarket.com/

11 http://www.wheredidmytaxgo.co.uk/

12 http://data.gov.uk/dataset/public_expenditure_statistical_analyses

13 http://www.airtext.info/

14 http://static.london.gov.uk/mayor/environment/air_quality/research/emissions-inventory.jsp

15 http://www.pitchup.com/

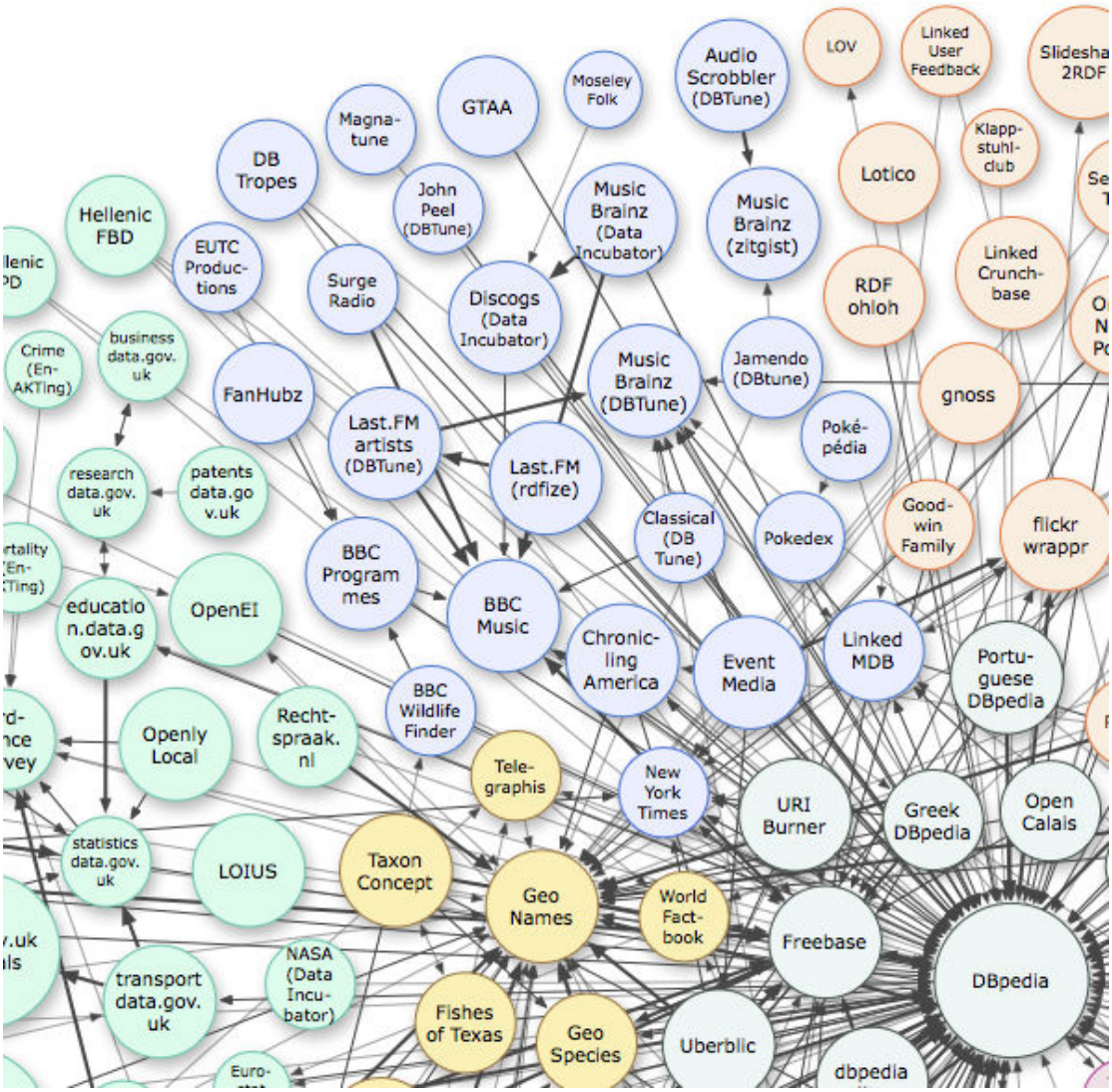
16 http://www.reegle.info/

17 http://lod-cloud.net/

18 http://www.bbc.co.uk/programmes

19 http://linkedmdb.org/

20 http://linkeddata.org/linked-data-shopping-list



A small section of the Linked Open Data Cloud as of September 2011

What might people do with the data held by classification boards? The question is unanswerable since it is up to individual developers and companies to create their applications. The release of open data is often accompanied by Hackathons[21] where developers are invited to turn up and see what they can do. Rewired State[22] is particularly good at organising these events.

Although we don't know for sure what people might do with the data, we can speculate.

Classification board data includes a lot of data besides the actual classification - that's useful for listings and review applications. It might also be mixed with other data to create things like maps showing film locations or stars' homes and if it's the classification board data that's being used, well, that consumer information is right there with it.

[21] http://open.nasa.gov/blog/tag/hackathon/

[22] http://rewiredstate.org/

Linked Data

As already noted, most applications that use open data are based on a single data set. The corollary of that is that most data sets are published as simple tables, almost always exported from Microsoft Excel in the non-proprietary format known as CSV. The showcase applications though use a format known as Linked Data in which data sets make links with related items in other data sets. That's what all those lines are in the LOD cloud diagram above - they're pointers from one data set to another so that, for instance, a reference to James Bond in one data set is linked to data about the same character in another. At the centre of the diagram is a data base derived from Wikipedia, known as DBpedia[23]. The creation of DBpedia was at the heart of the project that created the LOD diagram in the first place and was the catalyst for a lot of development in linked data. DBpedia takes a snapshot of Wikipedia every so often rather than using a 'live feed' but such has been the success of the project that Wikipedia itself is now working towards a (linked) data-centric architecture called Wikidata[24].

Linking classification board data to DBpedia would immediately increase the usefulness of the data. It is noteworthy in this regard that language-specific versions of DBpedia exist, that is, snapshots of the relevant language versions of Wikipedia including French and German (although not Dutch sadly).

There's no escaping the fact that publishing linked data does require a little more thought and effort than publishing CSV files but the value and attractiveness of the data increases as a result which, in the case of classification boards, means more classification data being presented directly to end users at the point of demand. The 5 Stars of Linked Data[25] developed by Tim Berners-Lee show the progression from simple publication through to linked data.

★	Available on the web (whatever format) but with an open licence, to be Open Data
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

This proposal tries hard to avoid technical detail except where unavoidable… and it is unavoidable briefly to explain the terms RDF and SPARQL. RDF is the technology used to publish linked data, SPARQL is its query language. A SPARQL Endpoint is a service that accepts and processes queries, returning results from the associated RDF data.

The recently launched W3C Linked Data Platform Working Group[26], with members[27] including IBM, Oracle, BBC and the EBU are testament to the growing importance of the technology and its move to the enterprise market.

---

[23] http://dbpedia.org/

[24] http://en.wikipedia.org/wiki/Wikidata

[25] http://www.w3.org/DesignIssues/LinkedData.html

[26] http://www.w3.org/2012/ldp/

[27] http://www.w3.org/2000/09/dbwg/details?group=55082&public=1

Policy Matters

The proposal set out above skirts around and hints at a number of policy issues that need to be addressed.

Fear of Misuse

A common reaction to the idea of publishing open data is a fear that it will be misused. The honest answer is, yes, it might be. But - no more than it can be now. The following image took less than 2 minutes to create and doesn't use any open data.

A faked image of the front of Call of Duty Black Ops showing the PEGI 3+ rating (the actual PEGI rating is 18)

It is already possible to mis-represent classification data. Usually this is through negligence but occasionally can be through deliberate malice. Publishing the data will not increase this malpractice. On the contrary: making the data easily available from a secure environment both as a service and as data increases the likelihood of it being used properly. It also makes it easier for anyone interested, be it a parent or a developer, to find the 'ground truth' - the verified data itself.

Make it easier to do the right thing and people are less likely to do the wrong thing.

Licensing & Server Load

The 5 Stars of Linked Data and the preceding discussion highlight the importance of an open licence. The UK Government's licence for public sector information[28] is generally held up as an example of how to do it. The main provisions are as follows.

> You are free to:
> - copy, publish, distribute and transmit the Information;
> - adapt the Information;
> - exploit the Information commercially for example, by combining it with other Information, or by including it in your own product or application.
>
> You must, where you do any of the above:
> - acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence;
> - ensure that you do not use the Information in a way that suggests any official status or that the Information Provider endorses you or your use of the Information;
> - ensure that you do not mislead others or misrepresent the Information or its source;
> - ensure that your use of the Information does not breach the Data Protection Act 1998 or the Privacy and Electronic Communications (EC Directive) Regulations 2003.
>
> *Note that this is an **edited and incomplete** version of the text and it must not be taken as authoritative.*

Does publishing data under an open licence like this not imply supporting a service with unlimited bandwidth and server capacity? Not necessarily. A common way to retain a degree of control over server load and, ultimately, the ability to pull the plug on a given service user is to require developers to register and include a token with each request. The Ordnance Survey does this for example with their Developer Licence[29]. But the greater the restrictions, whether legal or practical, the less the data will be used.

Ratings Apply to Specific Versions

The example API call above suggests that just 3 pieces of data are required to uniquely identify a film, game or TV programme. This is not the case, of course, as

[28] http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

[29] http://www.ordnancesurvey.co.uk/oswebsite/products/try-now/developers.html

many works have many different expressions, edits etc. The problem is that the industry reality doesn't match users' perceptions and developers are a lot closer to users than people in the industry. Therefore I suggest that to be successful, it will be necessary to include some sort of algorithm and for the service to return what amounts to a 'most likely' response - which may include multiple results. For example, something like this:

```
includeRating('board:BBFC', 'title:Titanic', 'director:Cameron');
```

is very ambiguous and returns 19 results (based on entering those terms in the BBFC search tool), but it might be possible to introduce some default filters, such as not including trailers, DVD extras etc. The API will need to support extra fields such that a query thike this would be supported:

```
includeRating('board:BBFC', 'title:Titanic', 'director:Cameron',
'medium:film');
```

The addition of the media type means we're down to 4 possibilities which are either rated 12 (from 1997) or 12A (from the 2012 re-release). Since the director and three named stars are all the same, the differences between them are hard to spot programmatically so the system should probably return all 4 results - that's OK - an application should be built to cope with that and allow the end user to make sense of the data. Of course the more specific the request, the more specific the answer, so that:

```
includeRating('board:BBFC', 'id:AFF150453');
```

is very specific and would return a single result since this is the BBFC identifier for the original cinematic release of James Cameron's Titanic.

The Industry View

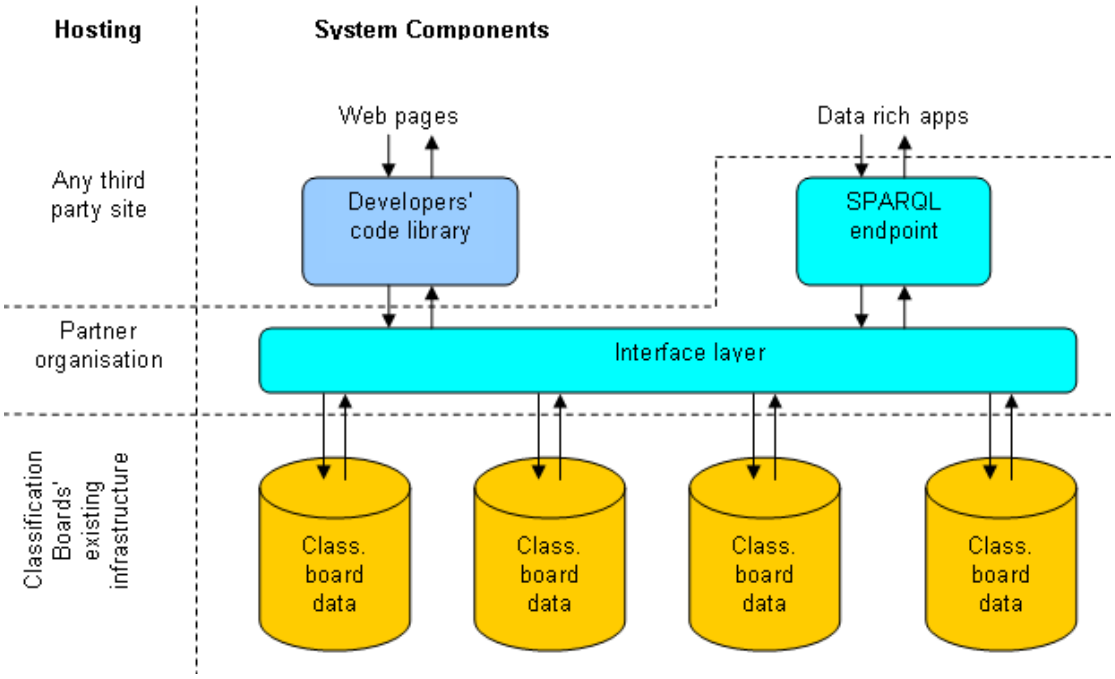
It was noted earlier that the there is a paucity of publicly available data concerning films, TV programmes and games. This is not because the data doesn't exist but because the relevant industries are reluctant to publish it. The open data philosophy does not sit well with the industries that are in the vanguard of the fight against copyright theft. And yet there are any number of Web sites devoted to these products that include a lot of the metadata (production companies, artists etc.) as well as trailers. Publishing classification board data adds to that, not the ability to make pirate copies, and, again, puts real data into the datasphere so that gaps are filled and false data can be less prominent.

Collaboration and Interoperability

The Linked Open Classification proposal assumes that multiple classification boards will use similar methods to make their data available. From a developer's point of view, that means that a single set of tools can be used to access any participating board's data. The one line API calls used in the examples above suggest that it is necessary to indicate which board's data is required - with no change in the way the data is requested or returned. That requires consensus on the infrastructure without requiring consensus on the classifications themselves.

In practical terms, implementing this proposal means adding a thin layer between a classification board's existing infrastructure and the outside world. Whether that thin

layer is a single service as the diagram below suggests, or is implemented separately by each board, is very much an open question.



A possible high-level architecture, with all details very much open to debate

Outline of a Pilot Study

The only real way to know what publishing classification board data really entails is to try it. My suggestion is that for a successful pilot study we need:
several classification boards;.
a research institute with relevant experience to help create the data model and build initial implementations.

As noted at the very beginning of this document, W3C is potentially interested too. It offers relevant expertise in data modelling but perhaps more importantly a neutral home for vocabularies and documentation, perhaps as either a Team or Member submission. For clarity, however, W3C standards can only be created by chartered working groups which in turn come about through community demand, notably from members.

The GAM consortium is a potential partner in a pilot study since this proposal is in line with their thinking and they may be a suitable partner organisation to host and manage the interpretation layer as part of their proposed infrastructure.

A possible sequence of events is:

- gathering of interested parties and agreement on terms of reference for the study and budget;
- outreach to other potentially interested parties such as ESRB and MPAA;
- engagement with the European Commission's Safer Internet Programme that may be able to offer some funding;
- agreement on a common vocabulary for describing classified works (almost certainly based on the FRBR[30] ideas of *work*, *expression* and *distribution*);

[30] http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm

- design and implementation of the interpretation layer, likely based on the RDB to RDF Mapping Language[31]) that allows a relational database to be accessed via a "virtual SPARQL query engine" - that is, to the outside world it would appear that classification data were being published as RDF even though in reality the data is stored in a table-based relational system;
- development of the minimal code library that allows developers to include classification data directly in their Web pages;
- development of a demo linked data application;
- a hackathon to elicit and explore further ideas;
- a report on the pilot with recommendations for follow up work.

Summary

The Linked Open Classification proposal offers:

- easy access to classification data at the point of demand;
- end users the classification systems with which they are already familiar and trust;
- technical interoperability without the need for harmonisation of classifications;
- an opportunity for the development of data-rich applications that promote games and films in new and innovative ways.

Phil Archer
2 August 2012
phil@philarcher.org

[31] http://www.w3.org/TR/r2rml/