From 8 miles high to ground zero: granularity in the information landscape

Gordon Dunsire

Presented at 14. seminar Arhivi, Knjižnice, Muzeji
17-19 Nov 2010, Poreć, Croatia

Published by Gordon Dunsire

Edinburgh 2012

From 8 miles high to ground zero: granularity in the information landscape

Gordon Dunsire
Presented at 14. seminar Arhivi, Knjižnice, Muzeji
17-19 Nov 2010, Poreć, Croatia

Abstract

The paper discuss the range of granularities that are encountered by users of the modern information landscape, from international, national, and organisational collections at the highest level, through records in union and local catalogues, indexes and finding-aids, to the metadata statement that is the low-level building block of the Semantic Web. Specific topics include interoperability between and within levels, and the contexts and environments appropriate to particular levels.

Information landscape

The phrase "information landscape" refers to the topology of an information environment. The components and interactions of information environments have been discussed at AKM in previous years (2005[1], 2006[2]]. Of particular interest is the idea of a common information environment which focuses on the user and removes barriers between archives, libraries, and museums to provide seamless, joined-up services and the widest range of resources supporting the information-processing needs of all users in a community.

A topology is a surface. The mountains and valleys of the information landscape indicate places where there is a high or low likelihood of a user finding information relevant to their needs. The topology of an information landscape is shaped by metadata, lumped together in records which are aggregated or clumped to form catalogues and other types of finding-aid. Records might be thin, using only a few metadata elements as in basic Dublin Core, or thick, with many more elements at a finer level of detail, as in the library domain's MARC format. Catalogues can describe information resources held in a single collection, or may be joined together to represent many collections held in one or more institutions. A single piece of metadata, for example stating the subject of a work, represents the finest level of granularity in a library landscape, with records, catalogues, and union catalogues forming increasingly larger aggregates of coarser granularity. Aggregations are formed in different ways in archive and museum landscapes, so the topology of a common information environment is likely to expose more levels of granularity than in a landscape bound to a single domain.

The basic environment is determined by the total aggregation of metadata it contains. This fixes the initial landscape that can be explored by a user. A search in an information environment typically filters

out most of the metadata irrelevant to a user's needs; what is left forms a new landscape specific to the user at that point in time. Real landscapes in the physical environment change too, but over very long periods of time as a result of geological processes; in a virtual information landscape, a mountain becomes a valley and a valley a mountain at the click of a button. The shape of the landscape is therefore determined by both subjective and objective factors. Each user has their own, self-directed requirements which result in a different landscape being presented by each search of the metadata. A search will usually aggregate the metadata into peaks formed by relevancy to the user's requirements; the precision of the search governs the height of a peak while recall determines the breadth of the base of the mountain. On the other hand, the granularity of the metadata records and aggregations is fixed, and determines the smoothness of every landscape that can be presented in the environment. Sparse metadata makes the peaks and troughs, the mountains and valleys, more fuzzy with less definition; richer metadata results in a more jagged appearance reflecting the detail of the information resources described. A high recall search on thin metadata results in the most rounded, amorphous landscapes, while a high precision search on rich metadata gives the most detailed landscapes. It is as if the former is a distant view of the latter, with the user being unable to get close enough to see the detail. The granularity of metadata affects the focal length of any view of an information landscape.

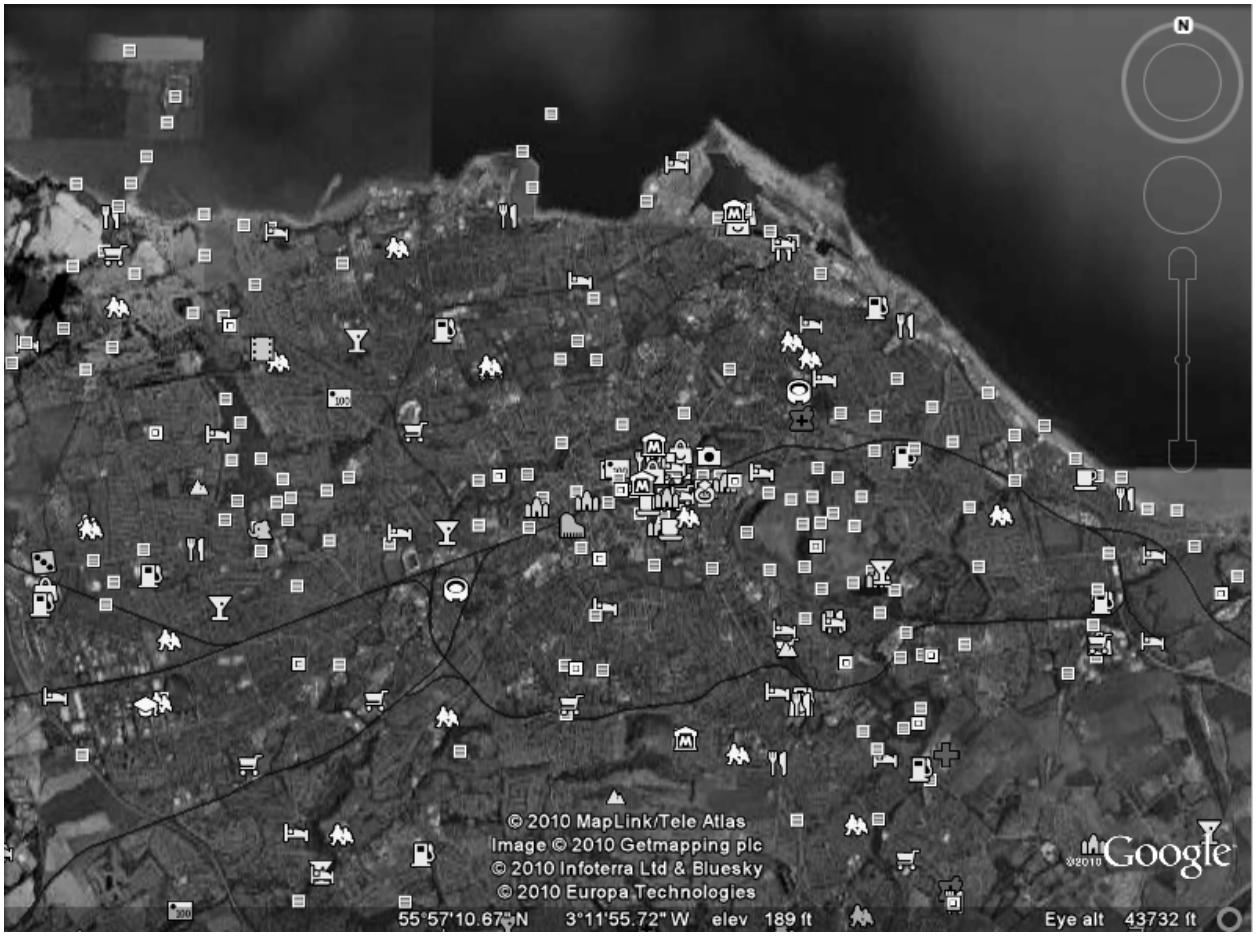
The metaphor of environments, topologies, and landscapes collides with reality where collections of physical resources are considered. Such resources are stored in actual places on the surface of planet Earth, so metadata about locations is required by users of information environments to allow them to select and obtain what they want. For example, the availability of a specific item to a specific user may be governed by the distance from the user to the item's location, local transportation services, etc. These metadata are the intersection between the metaphorical and actual landscapes of physical resources.

The least granular information landscape for physical resources consists of the inhabited regions of the Earth's surface, assuming that uninhabited places such as oceans and the frozen North and South polar regions do not contain physical information resources of any interest. This is not strictly true: there are quasi-museums in polar exploration camps, and underwater "zoos" of marine life; but generally-speaking resources are created, collected, and stored by humans in accessible places.

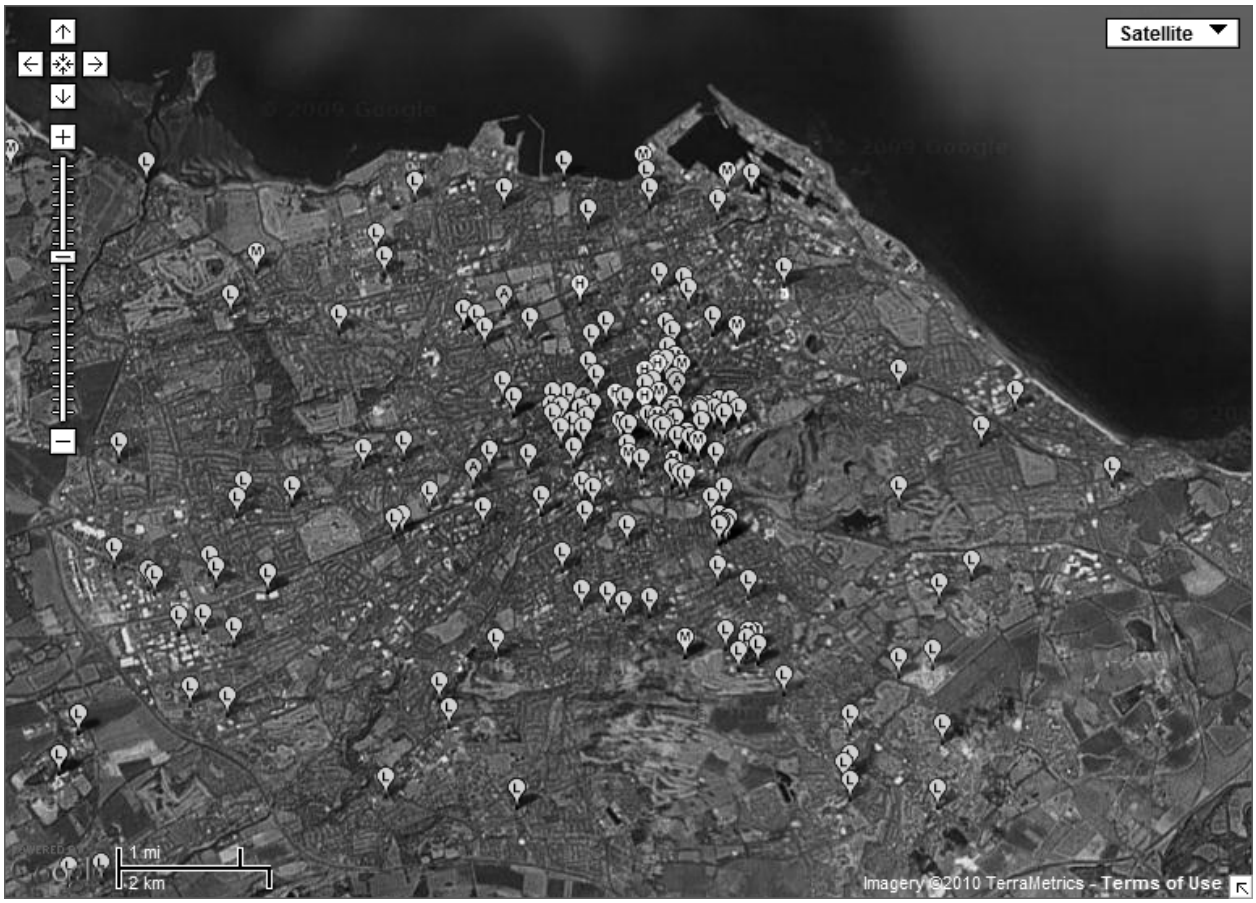
There is, as yet, no common information environment on a global scale for archives, libraries, and museums. The largest environments exist only at regional and national levels, although there have been, and continue to be, many initiatives to develop international aggregations of metadata for physical and digital resources and aggregations of digital resources themselves. There is, however, at least one attempt to create a global information environment for locating resources located in, or with the subject of, specific places, although it extends to information about resources of all kinds and not just those which are bibliographic, cultural, or heritage. This is Google Earth[3] and its close cousin Google Maps[4].

The Google Earth service emphasises the scale of its environment: the opening screen shows the planet as it would appear from over 68000 miles away in space. This is the distance required to conveniently display the whole hemisphere facing the user: the light side of the Earth. The only details that can be seen are the outlines of countries superimposed on the geophysical land masses. A user is able to rotate

the display to show the far side of the Earth, and zoom in on any point to display more detail. As more detail is revealed, the locations of resources appear, marked by icons. Each broad category of resource has its own icon. Resources include services for catering and accommodation such as restaurants and hotels, and places for recreation such as tourism, sport, and exercise. It is worth noting that these may be significant additional factors in determining a user's choice of cultural heritage information resources: is there a railway station within easy reach of the archive, and will I be able to have lunch nearby? Zooming in is the equivalent of reducing the distance from the user to the landscape. For example, the whole of the country of Scotland can be displayed on screen from a virtual distance of around 550 miles up. At this level of zoom, the locations of only a few resources are displayed. These actually represent clusters or aggregations of resources, and prevent the interface from becoming too cluttered to use. So a single icon is displayed over a city which actually contains many, many resources. Zooming further in reveals more and more locations. At eight miles high, or just over 42200 feet, the whole of the Edinburgh area is displayed, marked with hundreds of locations, as shown in Screen 1. Only a few archives, libraries, or museums are shown.



Screen 1: Google Earth view of the Edinburgh area from approximately 8 miles high.



Screen 2: Scotland's Information view of the Edinburgh area from approximately 8 miles high, using Google Maps.

There are no technical barriers to including the locations of all archives, libraries, and museums on the same display. At this level of zoom, the satellite imagery used by Google Earth is the same as that available in Google Maps, and it is possible to add Maps markers to the Earth service. The Scotland's Information service[5] is a mash-up between the collection-level description metadata held in the Scottish Collections Network (SCONE) database and Google Maps. The metadata include latitude and longitude for geolocation of the archive, library, and museum buildings holding collections. The Scotland's Information location markers are not clustered at different levels of zoom, as happens in Google Earth, but it is worth noting that there are as many of them in the Edinburgh area of the Scotland's Information landscape as there are in Google Earth from the eight-miles high perspective, as shown in Screen 2. Although the Google Earth service reveals more markers at closer zoom levels, the Scotland's Information service provides access to finer granularity through additional metadata from the collection-level descriptions. Every location marker is associated with an institutional collection for that archive, library, or museum, covering all of the resources held at the location. Each institutional collection may be an aggregation of sub-collections which themselves are aggregations of sub-sub-collections, and so on. For example, the John Donaldson collection of musical instruments[6] is part of the

Edinburgh University collection of historical musical instruments located within the Reid Concert Hall Museum of Instruments, which is a building marked on the map.

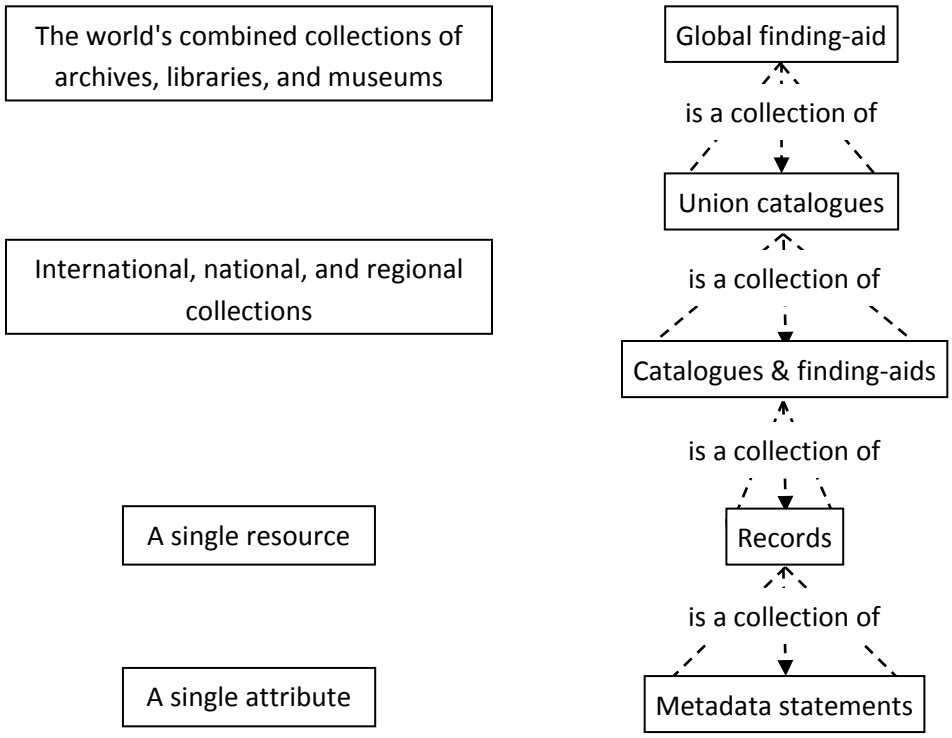


Figure 1: Levels of granularity, from the collection of the world's archives, libraries, and museums, to a single attribute of a single resource.

Furthermore, the collection-level metadata include links to online catalogues, exposing the finer granularity of a single catalogue, its records, and their component metadata. Digital images of some of the instruments in the John Donaldson collection are available online, and can be found by browsing down through a multi-level finding-aid. This example therefore demonstrates the potential for a seamless descent through multiple layers of resource granularity: starting from a global collection of all information resources and moving through international and national collections to institutional collections, then to special collections, categories within them, and eventually a single item. The granularity of these physical and virtual aggregations of resources is mirrored by aggregations of the corresponding metadata, as shown in Figure 1. A global common information environment of cultural heritage organisations which supports user-generated landscapes at all levels is possible, but a long way from instantiation. Nevertheless, we can see large-scale metadata aggregations like OCLC's WorldCat[7] as significant progress. WorldCat offers the traditional facilities for landscape generation by a user searching its metadata records, which are mostly from the library domain. The user is also able to identify the locations of a chosen physical resource on Google Maps, but only one location at a time. The user chooses a location before seeing it on the map, rather than seeing the map before choosing a location.

Maps are just one way of displaying an information landscape. Other graphical methods can be used for landscapes which are not focussed on the locations of physical resources. For example, tag clouds can display levels of subject granularity, as in the Dewey Browser[8], where the size of a caption (tag) indicates the number of resources classified by the corresponding notation.

Whatever method is used to render and visualize the topography, the granularity of the user-generated landscape is determined by the granularity of the metadata in the environment. And the finest level, what might be called the atomic level, of granularity currently used in cultural heritage information environments is the record, not the component metadata element. The record displayed to the user is essentially indivisible. It may not be all of the metadata stored in the environment, if administrative and technical elements are concealed from the end-user, but what is displayed is a unitary lump which often contains elements irrelevant to the user's specific requirements. This can be contrasted with Web search engines where results are displayed in a way that contextualises the search term in relation to the content of the resource itself: the resource is the record. For example, Google shows search results as extracts from the documents which contain the terms input by the user, representing a finer level of granularity than that of the metadata record. This is not to say that current Web search engines are solutions to developing common information environments suitable for cultural heritage collections and their users, even if the obvious problem of non-availability of offline content in physical resources is ignored. The environment of these search engines is the Web of documents, and landscapes are generated with extremely simple metadata structures. Essentially only one metadata attribute is used. It can be expressed as "contains terms", and a typical metadata statement is structured as "The document located at this URL contains terms [terms input by user]". These are very thin metadata and are inadequate to support detailed landscapes; for example "This document contains terms 'bagpipes'" is true for a resource which states "this resource is not about bagpipes". What should be a valley becomes a mountain. This contrasts with standard archive, library, and museum metadata where hundreds of different attributes have been developed in response to users' needs. Examples of typical metadata statements in a common information environment are "This book with identifier Y has subject 'Bagpipes', "This object with identifier Z is made of goat skin", "This set of letters with identifier X has author 'G. MacDonald', etc.

This range and variety of attributes is used to construct very rich metadata records, but their full potential for revealing the finest detail of the common information environment will only be realised if there is a way to refine the atomic level of granularity from the record to the individual metadata element or statement. The user would then be able to define and navigate the environment in a more seamless, flexible way, while the lumpiness of resulting landscapes could be reduced by displaying the "records" required for resource identification and selection as sets of individual metadata statements customised to meet the user's specific needs. That is, the record is a dynamic set of those metadata relevant to the user-generated landscape.

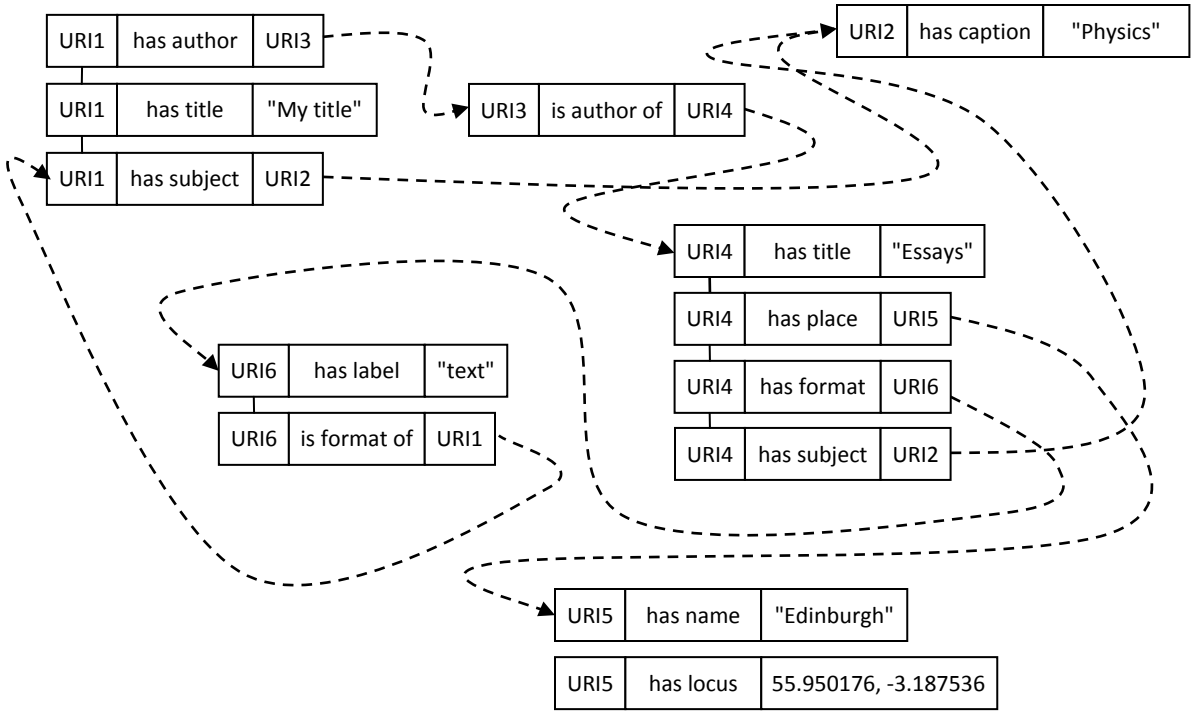


Figure 2: RDF triples linked by URIs to create linked data.

The resource description framework (RDF)[9] of the Semantic Web provides a technology to do this. Metadata are represented as single, three-part statements known as triples which can be connected using uniform resource identifiers or URIs[10]. Using the same URI for the first part of a set of triples allows the set to be brought together as a "record"; using the same URI for the third part of one triple and the first part of another allows the two triples to be connected in a chain of so-called linked data for navigation and resource discovery. A very simple example is shown in Figure 2. RDF is therefore suitable for adding the finest useful level of granularity to a common information environment and refining the landscapes it can support. The Semantic Web as a web of metadata complementing the Web of documents has parallels with the aggregation of archive, library, and museum metadata to support the common information environment. Indeed, in many ways the Semantic Web is, or subsumes, the common information environment.

Ground zero, the closest a user can get to the surface of an information landscape, is therefore an individual metadata statement or RDF triple. If archives, libraries, and museums can convert their legacy metadata of hundreds of millions of records to generate billions of triples, the result will be a critical mass of linked data that ensures seamless connection between atomic metadata statements, guaranteeing a "chain reaction" or landscape to meet every user's needs. The beginning of a critical mass for the Semantic Web in general is shown by the linking open data "cloud"[11]. Archives, libraries, and museums can contribute not just large quantities of metadata, but also high quality content resulting from standards developed over many years of professional activity. And combining this

metadata with modern graphical displays will provide a landscape suitable for every user; a common information environment for the common man.

References

[1] Dunsire, Gordon. The common information environment: a newly emerged concept. In: 9th Seminar on Archives, Libraries, Museums , 23-25 Nov 2005, Poreč, Croatia. Original English expression available at: http://cdlr.strath.ac.uk/pubs/dunsireg/akm2005cie.pdf
[2] Dunsire, Gordon. Future information environments: deserts, jungles or parks? In: 10th Seminar on Archives, Libraries, Museums, 22-24 Nov 2006, Poreč, Croatia. Original English expression available at: http://strathprints.strath.ac.uk/6028/1/akm2006futures.pdf
[3] Google Earth. Available at: http://www.google.com/earth/
[4] Google Maps. Available at: http://maps.google.com/
[5] Scotland's Information. Available at: http://www.scotlandsinformation.com
[6] Scotland's Information: John Donaldson collection of musical instruments. Available at: http://www.scotlandsinformation.com/SIColnShow.cfm?MC=sca,typ,ove,pzl&CI=5962
[7] WorldCat. Available at: http://www.worldcat.org/
[8] DeweyBrowser. Beta v2.0. Available at: http://deweybrowser.oclc.org/ddcbrowser2/
[9] Resource description framework (RDF). Available at: http://www.w3.org/RDF/
[10] Uniform resource identifier (URI): generic syntax. Available at: http://tools.ietf.org/html/rfc3986
[11] The Linking Open Data cloud diagram. Available at: http://richard.cyganiak.de/2007/10/lod/