Future information environments: deserts, jungles or parks?

Gordon Dunsire

Presented at Archives, Libraries, Museums 10 (AKM10), Porec, Croatia, 2006

Published: Glasgow, Scotland : Centre for Digital Library Research, 2008

Title

Future information environments: deserts, jungles or parks?

Presented at Archives, Libraries, Museums 10 (AKM10), Porec, Croatia, 2006

Author

Gordon Dunsire

Summary

This paper discusses the basic functions of a common information environment and how they are supported by metadata. Several distinct categories of information landscapes are described, characterised by the availability and quality of metadata at the item and collection level. The paper suggests elements of professional practice which can improve the functionality of landscapes, and presents an illustrative scenario of how a common information environment might be effective.

Keywords

common information environment; information landscapes; metadata aggregation

Future information environments: deserts, jungles or parks?

This paper is a follow-up to the paper presented last year at AKM9 {1}. That presentation was summed-up with "A common information environment offers easy, convenient access to the widest range of information resources catering for the widest range of needs of the widest range of users - It requires the collaboration of archives, libraries, museums and governments" {2}. This paper focuses on the general functionality of resource discovery and the metadata infrastructure required to support it within a common information environment and discusses the impact of current developments in that infrastructure.

The metaphor of an environment can be readily extended to provide one-word labels for complex processes. An example is the term "landscape".

The term is associated with the use of metadata describing collections of information resources. Such metadata fall into two categories: metadata about the collection as a whole, and metadata about its constituent parts, the items in the collection. Collection-level metadata are used to find and identify collections which are likely to meet the specific information needs of the user, and to identify finding aids consisting of item-level metadata for the resources in those collections. The set of collections and associated finding aids is called a "landscape". In a common information environment, a landscaping service must have access to metadata about collections and sub-collections held in archives, libraries, museums, and other information services. The metadata must also be rich enough to support a wide variety of user needs, including, amongst others, the identification of collections by subject topic, location, accessibility, educational level, resource format, and curatorial management. A landscaping service must also be able to exploit all of the relevant item-level finding aids in the environment, including those based on aggregations of metadata from multiple sources. Aggregations such union catalogues, multiple archive repositories and other community-wide finding aids may have different public interface styles and functions from local finding aids such as online catalogues and archival descriptions. Such choice can improve the range of landscapes that can be supported. Coverage of aggregated item-level metadata in a community-based service may overlap wholly, partially, or not at all with any specific local service. A large-scale information environment will cover multiple communities, and aggregations of aggregations with equally varying coverage may become available. This results in a wider range and choice of item-level finding aids; a good landscaping service should be able to guide the user to the best set of finding aids which suits their requirements for coverage, ease-of-use, and functionality.

Landscaping also requires aggregation of collection-level and item-level metadata because aggregation increases coverage and improves efficiency in searching. The online availability of local aggregations of item-level metadata has been, and remains, a driving influence on the development of information environments. Although coverage is a major factor, quality has a significant role to play, and the

effectiveness of an information environment will be determined by the coherency, consistency, and completeness of its aggregated metadata.

One ideal feature of a landscape is that it should be transparent to the user. The professional and technical complications of collection versus item description and metadata format, content and aggregation should not be allowed to adversely affect the user's interaction with the environment; their experience should be as seamless as possible. The user should be able to tell the service what topic they interested in, whether a person, organisation, place, event, or subject, and the service should display information about resources relevant to the topic. Such information should include details of the items, where they are located, when they can be accessed, and how they can be used. The user should then be able to make further selection of resources based on this data, change or broaden their search topic, and combine searches.

Stretching the environmental metaphor still further, this paper now uses the term "ecology" to describe a basic type of information environment. An information ecology is characterised by the ability of its metadata, at all levels of granularity of aggregation, to support landscapes hospitable to the user. Like natural ecologies, information ecologies may change over time, it is possible for one to evolve into another, and an information environment may contain multiple ecologies at different granularities.

Most current information environments are based on communities of professional practice at the sectoral level within a general domain: research libraries, business archives, town museums. Item-level metadata is consequently focussed on the needs of the local community, especially with regard to content. The choice of format is often governed by issues arising at the higher domain-level, such as collaborative cataloguing services or the availability of technical support and advice. Indeed, many of the available item-level metadata aggregations are based on format, and therefore professional domains. Cross-domain aggregations are rare and apparently expensive to maintain, as they are often confined to a small number of sources, or are developed within well-funded communities. Item-level and collection-level metadata consistency and completeness also varies widely and wildly between and within sectors. Aggregations of collection-level metadata are sparse, often little more than domain-based directories. As a result, common information environments which cover all domains and sectors are currently capable of supporting only the barest of landscapes. They are deserts.

A desert landscape is the same in every direction, with few salient points. The occasional hint of green may be a rich oasis of resources, an isolated collection, or a discarded website masquerading as an empty soft drink bottle. It's difficult to tell without spending some time getting closer. The information prospector has a long and tiring task in front of them, even with the aid of a treasure map. The path will be impeded by the bleached bones of unfinished digitisation projects, accompanied by error messages from the local finding aid flapping in the desert wind. All (data)

mines encountered will have to be examined to have any chance of finding the mother-lode, but many will have been abandoned years ago. There will be little or no help available on the way, and the prospector may never find what they seek.

Many current common information environments contain the equivalent of oases, well-developed and rich resource discovery services based on metadata aggregation but with a relatively small scope. It is possible that these may become the seed for expansion, to enrich the functionality of the whole of the environment. There are several developing technologies which can act as growth-accelerators for aggregation. The oldest method of aggregating item-level metadata is exemplified by the physical union library catalogue. Each member of the community sends, or pushes, a copy of its metadata to a central repository. This activity is repeated as required to keep the metadata copy synchronised and up-to-date with the local version. In general, both the local and central finding aid will display the same item-level metadata, although the functionality of the search interfaces may be quite different. Another method of creating physical aggregations is "harvesting", where a copy of the local metadata record is taken, or pulled, by an automated process controlled by the central repository. Each member of the community allows its local metadata to be copied in this way, but may have little or no control over what use is made of the central repository. Although it is technically possible for a community to restrict which services can harvest its metadata, in practice the software and standards allowing low-cost and rapid deployment of this aggregation method have been developed by open access communities. As a result, metadata complying with the most common protocol, the Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH), can be aggregated by anyone with readily available free software. It can also be amended by the aggregator, and re-presented for further harvesting by other aggregators. It is possible, and therefore probable, that a high-level aggregation service might harvest metadata for the same item from its local repository, and from one or more intermediate "central" repositories.

Virtual metadata aggregation services search metadata in situ in multiple local databases and only aggregate it temporarily for display to the current user; this is also known as meta-searching. Developments in relevant technologies such as SRU/SRW and improved interoperability between physical and virtual aggregation systems are supporting the evolution of hybrid services. The local databases used in meta-searching may themselves may be physical aggregations from multiple sources, so duplicate metadata will appear in virtual aggregations too. Push-type physical aggregation and virtual aggregation methods result in item-level metadata which increases duplication at the Work, Expression and Manifestation, but not Item, levels described by Functional requirements of bibliographic records (FRBR); the FRBR Item describes the local copy of a resource, and it is unlikely that this metadata will have been duplicated in another local database. Pull-type physical aggregation can duplicate metadata at the FRBR Item level as well, duplicating the description of the same copy of a resource. Commercial services are also becoming more interested in metadata aggregation, which may in turn

decrease the objectivity of descriptive records; some duplicates of metadata may be more reliable than others.

Common information environments are therefore increasingly likely to contain multiple copies, many times over, of item-level metadata. Some records may refer to separate instances or copies of a resource distributed in different collections; some records may refer to the same instance of a resource, such as a manuscript or digital information object. Some records for the same resource will be identical in descriptive content; other records for the same resource may vary in content. Records from different communities describing different types of resources are not likely to share the terminologies used to describe people, organisations, places, events and subject; the same thing will be described differently. Common information environments using every metadata aggregation which becomes available will support access to a richer variety of resources, but the landscapes they generate will inevitably be lush with duplicate metadata and multiple finding aids. They will be equally dense and overwhelming in every direction, with little chance of seeing the wood for the trees; the landscapes will be jungles.

The information prospector in this ecology will then to have to hack a way through a mass of duplication, constantly distracted by what looks like something new which turns out to be something already discovered. Prominent and attractive features will beckon, only to emit a bad smell when close-up. Many paths, laboriously traversed, will lead the prospector back to the starting point. The general dampness and stickiness of the landscape will evoke fond memories of when it was a desert.

Neither of these extreme ecologies is capable of delivering the promise of a good common information environment. Some kind of balance is required, a cultivated information ecology.

The metadata required to support this must be as consistent, coherent and complete as possible, to support the removal of duplicate metadata from landscapes. The use of machine-readable identifiers, such as standard numbers and codes, is of limited applicability for de-duplication because of the coverage and compatibility of the various identification systems employed by different communities. Metadata aggregated from different sources should be contextualised, appropriate, and relevant, to help the user to make sense of the landscape. Metadata must be clear and distinguishable, to allow the landscape to include a network of discovery routes. The metadata must also be rich enough to be used flexibly by the landscaping service to support as wide a range of user needs as possible.

The resulting landscapes will appear to follow an orderly plan. Some may contain patches where things are growing wild, while some may be completely formal down to the smallest detail, but all will have clear signposts and paths to the major sections. The information prospector can leave the water bottles, faded treasure

maps, machetes and mosquito nets at the entrance, for behind it lies a park or garden waiting to be enjoyed.

In a final extension of the environmental metaphor, it is clear that is we, the professionals working in archives, libraries, museums and other information management services, who are the gardeners. And we have many tools and techniques at our disposal. These include the adoption of authoritative headings, or consistent and coherent labels for referencing cultural information resources, covering people, organisations, places, events, themes and topics. A variety of quality control procedures and workflows can ensure currency and completeness of metadata and avoid duplication or make it easy to ignore if it occurs. Full-text indexing and smart tagging technologies are available to enrich the metadata for digital and digitized resources. The digitization of secondary finding aids such as abstracts, summaries and indexes, and tertiary aids such as directories can help to extend the coverage and granularity of the environment, allowing it to support larger and more varied landscapes.

There has been a lot of metaphor in this paper, so it seems appropriate to end with a return to earth, in the form of a speculative glimpse at a landscape from the near future, generated by a cultivated information ecology:

The user is physically located at the top of Edinburgh's High Street, also known as the Royal Mile. Edinburgh Castle has just been visited, and the user starts to walk down the street towards Holyrood Palace. The user has the latest gadget: an i-PIE, or integrated personal information environment. It is a portable tablet computer with built-in media processing, telephone, and high-definition still and video camera. It is connected to a free city-wide broadband wireless network. It knows where the user is, using a gps decoder, and the user has given it a profile of their general interests. After the user has walked several metres, the i-PIE buzzes …

The i-PIE informs the user that they are opposite a commemorative plaque set into the wall, displays a photograph of the plaque, and a clear transcription which reads "In a house on the east side of this close Robert Burns lived during his first visit to Edinburgh 1756". There are links to further online information about the current building, the original house, "closes" (narrow alleyways between buildings), Burns' visit, and Edinburgh in 1756. There is also a list of physical collections located within a specific distance of the user (1000 metres) which contain items relevant to Robert Burns and Edinburgh. These include: an exhibition of Burns' manuscripts at the National Library of Scotland; the Edinburgh Room at the central public library; an exhibit on farming in eighteenth-century Scotland at the Royal Museum of Scotland; a permanent display of Burns' signatures and entries in excise (customs) records at the National Archives of Scotland. Against each entry is displayed a series of symbols which tell the user at a glance whether there is easy physical access to the collection, its opening times and entrance fees, and the availability of finding aids. Other symbols lead to map-based, textual or audio directions to the location, or to nearby restaurants, bars, theatres, public transport; the user is a

tourist, and finds all of this information useful. So do the local information professionals, businesses and government organizations which supply and maintain the metadata, because tourism is a major source of income for Edinburgh and Scotland as a whole.

References

{1} Zajedničko informacijsko okruženje: pojava novog koncepta = The Common Information Environment: a newly emerged concept", conference proceedings / Gordon Dunsire. In: 9. seminar Arhivi, knjižnice, muzeji: Mogucnosti suradnje u okruženju globalne informacijske infrastructure. – Zagreb, Hrvatsko knjižničarsko društvo, 2006.
Available in English at: http://cdlr.strath.ac.uk/pubs/dunsireg/akm2005cie.pdf

{2} The Common Information Environment: a newly emerged concept / Gordon Dunsire. Available at: http://cdlr.strath.ac.uk/pubs/dunsireg/akm2005cie.pps