The Knowledge Engineering Review, Vol. 30:2, 220–236. © Cambridge University Press, 2015
doi:10.1017/S0269888914000319

Distributed aggregation of heterogeneous Web-based Fine Art Information: enabling multi-source accessibility and curation

FRANCES BUCHANAN[1], NICCOLO CAPANNI[1] and HORACIO GONZÁLEZ-VÉLEZ[2]

[1]*School of Computing, Robert Gordon University, St Andrew Street, Aberdeen AB25 1HG, UK;*
e-mail: fabuchanan@lumison.co.uk; n.capanni1@rgu.ac.uk;
[2]*Cloud Competency Centre, National College of Ireland, Mayor Street-IFSC, Dublin 1, Ireland;*
e-mail: horacio@ncirl.ie

Abstract

The sources of information on the Web relating to Fine Art and in particular to Fine Artists are numerous, heterogeneous and distributed. Data relating to the biographies of an artist, images of their artworks, location of the artworks and exhibition reviews invariably reside in distinct and seemingly unrelated, or at least unlinked, sources. While communication and exchange exists, there is a great deal of independence between major repositories, such as museum, often owing to their ownership or heritage. This increases the individuality in the repository's own processes and dissemination. It is currently necessary to browse through numerous different websites to obtain information about any one artist, and at this time there is little aggregation of Fine Art Information. This is in contrast to the domain of books and music, where the aggregation and re-grouping of information (usually by author or artist/band name) has become the norm. A Museum API (Application Programming Interface), however, is a tool that can facilitate a similar information service for the domain of Fine Art, by allowing the retrieval and aggregation of Web-based Fine Art Information, whilst at the same time increasing public access to the content of a museum's collection. In this paper, we present the case for a pragmatic solution to the problems of heterogeneity and distribution of Fine Art Data and this is the first step towards the comprehensive re-presentation of Fine Art Information in a more 'artist-centric' way, via accessible Web applications. This paper examines the domain of Fine Art Information on the Web, putting forward the case for more Web services such as generic Museum APIs, highlighting this via a prototype Web application known as the *ArtBridge*. The generic Museum API is the standardisation mechanism to enable interfacing with specific Museum APIs.

1 Introduction

A huge proportion of the adult population in the world now has access to digital technology and the Internet. This access brings with it the power to not only consume information but also the ability to publish it. Individuals, referred to by Shirky (2010: 64) as the 'people formerly known as the audience', have gone from merely consuming information in front of a television to actively contributing, creating and sharing all forms of digital media content. The current population has grown up with the Web and is adapting to its changes. Each new generation grows up familiar with the Web as it is *in their time* so that the growing interactivity becomes second nature. The infrastructure that facilitates this is the World Wide Web, commonly designated 'the Web'. Although this was initially a repository of interlinked hypertext documents written by a small proportion of the Web population, the majority of Web users had only passive access to browse or read. This has changed considerably, and in respect of this paper, in three important ways.

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

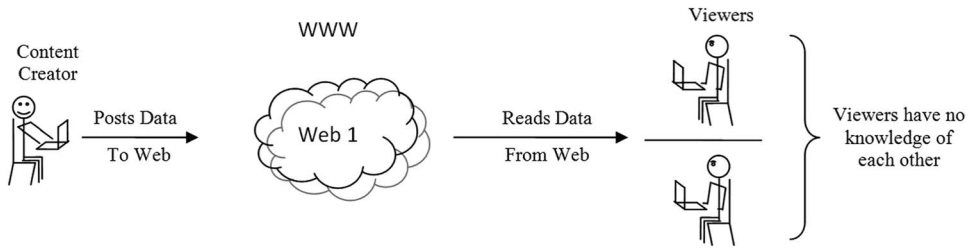

Figure 1 Users are unaware of any other users

First, hypertext has been enhanced with images, videos, multimedia applications and documents in various formats. Second, the introduction of Web applications has enabled information contribution on the same scale as information browsing; and third, Web services have opened up previously repository-centric data in a way that encourages others to analyse and augment that data. This change of approach to the Web is often referred to as 'Web 2.0' and in human terms is akin to a library being transformed into a community of authors (O'Reilly, 2007).

A key component to the developing Web is users' awareness of each other. Previously, a user browsing a Web resource would be unaware of any other users (see Figure 1). As participation increased so did awareness, through time filtered posting up to 'instant' communication, as shown in Figure 2. The tools of this new community range from the semi-individualist blogs, file hosting/sharing services 'torrents', through to highly interactive media sharing sites, multi-purpose Web applications, and the more communal wikis, mashups and social networks (see Figure 3). These tools create multiple sources of data, its associated meta-data as well as interpretations, contradiction and outright conflict in relation to the original materials. Crucial to permitting contribution is that anyone with access to the Internet is able to publish and distribute digital information for minimal effort and cost.

Typically referred to as *open data*, the open source style contribution to Web data where access is (largely) without restrictions from licensing, patents or copyright has resulted in natural virtual groupings of individuals with the most diverse commonalities. Websites are taking on a more two-way conversational and interactive role, facilitating the sharing of information, the establishment of communities of people with similar interests, and the creation of opportunities to comment and contribute.

Such a culture of openness of information and data is spreading. There is a growing list of organisations that are opening up their previously repository-centric art-related data with a view to increasing transparency and dissemination, whilst encouraging others to analyse and add to that data. The list includes government agencies, the BBC, *New York Times*, *The Guardian*, and several universities, museums and archives.

Of singular importance to this paper are the Web services providing Fine Art open data. The provision of programmable Web access to a museum's archive of information is seen by these organisations as a new means of increasing the exposure of their collections whilst creating a digital dialogue with developers and the wider community. In the context of the present paper, it is seen as an opportunity to aggregate Fine Art Information in a more 'artist-centric', and therefore user-friendly, way.

This paper seeks to examine the place of a Museum API (Application Program Interface) in the trend towards a more comprehensive aggregation of Fine Art Information. It begins, in Section 2, by taking a general look at the current state of Fine Art Information, highlighting the problems inherent therein, and describes the contributions of this project. It significantly extends our initial work (Buchanan *et al.*, 2011) by reporting the introduction of the carefully designed user interface for ArtBridge as well as providing a holistic analysis of the application of large-scale Web-based systems to Fine Art.

Section 3 discusses the method by which a solution can be provided, examining the issues limiting a centralised approach and favouring an open data approach. It then examines the transition from data to information via a review of a number of different Web applications that have been created using open data sources. In particular it highlights the way in which technology has improved the accessibility and quality of information currently available in the book and music domains. Examples from book and music domains can give valuable lessons in usability.

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

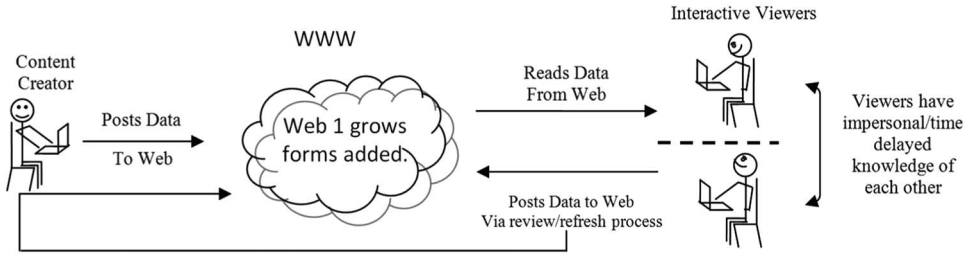

Figure 2 Users are semi-aware through time-filtered posting up to 'instant' communications

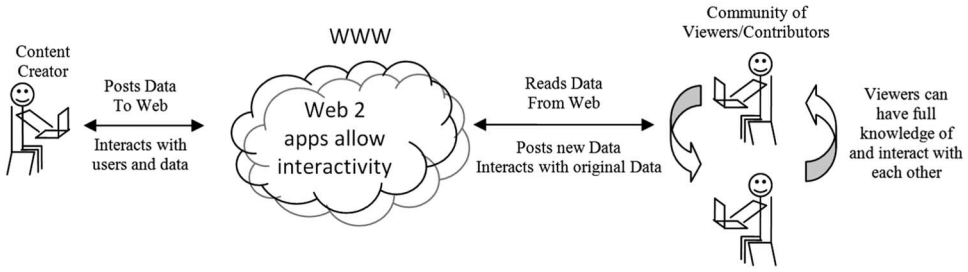

Figure 3 Interactive users through communal wikis, mashups and social networks

Section 4 presents ArtBridge, a Web application built using open data obtained via a number of Museum APIs.

In Section 5, the technical implementation of ArtBridge is given and the case is made for the further aggregation of Fine Art Information and therefore the need for the provision of more Museum APIs. Finally, Section 7 adduces the findings for our research to validate our data-centric hypothesis for the design of Web-based Fine Art repositories and APIs.

2 Background

Provenance is well understood in Fine Art and curators have typically got a well-established set of processes to determine the origins of a given artwork. However, such processes are not necessarily transferable, particularly when it comes to determine the necessary 'meta-data' to define electronic catalogues and collections in distinct museums.

Crawlers and search engines rely on the integrity of the data they are retrieving. Some progress has been made on this with the introduction of trust-based applications such as the 'Web of trust' (Artz & Gil, 2007), which is a community-based tool that offers feedback from user reviews and rating to increase the confidence in good data sources and reject ones that are poorly managed or actively destructive. This Web relies on human activity, after all there are millions of participants. It is open to abuse but the community contribution acts as a self-correcting mechanism. An alternative approach is to introduce Artificial Intelligence in the form of machine learning to improve the relevance of the information retrieved (Snásel *et al.*, 2009).

The Web is continually growing in data content and old data is often amended, replaced or deleted. The result is that the semi-intelligent software agents, which feed the search engines, usually referred to as Web crawlers, have an increasingly difficult task in gathering new data and confirming the relevance of previously indexed data. In short the Web is growing and changing faster than the indexing systems can keep up. Web crawlers and related approaches are currently incapable of full information gathering on the Web (Baeza-Yates, 2003) and building efficient meta-search engines remains a colossal endeavour (Meng *et al.*, 2002). Crawler technology is of course also improving but there are restrictions on them from the Web hosters' point of view. Anti intrusion, to prevent illegal access to resources, and subscription only data mean that a portion of the Web will remain out of crawler reach for the foreseeable future (Henzinger, 2001).

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

Hence, provenance-based retrieval of electronic data remains an open problem in computer science (Moreau *et al.*, 2008). While semantic search tools (Uren *et al.*, 2007) and ontology-based retrieval (Mayfield, 2002) have long been considered a suitable alternative, the indexing systems that currently dominate Web information retrieval are the generic search engines. These try to be all things to all surfers.

Within the museographic domain, generic search engines are often used to implement an institute's own search facilities. Although this allows specific tailoring of the indexing system, it has two drawbacks. First, the tailoring is customised to the specific institute, so inter-institute online cataloguing, curation and, in general, cooperation are tied to bilateral agreements. Second, the data indexed is restricted to the implementing institute so it does not allow the indexing of cross-institute data. The problem is akin to every museum independently building a catalogue system and API.

The Dublin Core (Weibel, 1997) has long been considered a suitable alternative to homogenise metadata and associated schemas to enable mapping of disparate cataloguing sources, without the requirement for a centralised data store. In theory, data can be widely dispersed across the Web in both location and format. Nevertheless, the result of the different mechanism for information retrieval is that specific rather than general APIs are being constructed for different organisations to index the same type of data. The BBC (Kobilarov *et al.*, 2009), both curate data sets from external sources and allow access to their own content via an API. This data is now subject to the schema of the BBC API and this may result in incompatibility to indexing with other APIs that follow a different scheme to the BBC.

A small number of specialised collaborative projects have specifically been funded that begin to demonstrate the possibilities for the aggregation of Fine Art Information. For example, the Google Art Project `www.googleartproject.com` is a website that brings together selected data from 17 different public galleries (Proskine, 2006). Each of those galleries has released high-resolution images of a selected group of artworks as a means to publicise both the content of their collections, and their physical gallery spaces (using the Google StreetView technology). It is not possible, however, to search this site by artist's name—rather the main purpose of the site appears to be to provide highlights from each gallery's collection. The information is gallery centric rather than artist centric, and indeed it includes only a small proportion of the artworks in each institution.

Then there is Culture24 `www.culture24.org.uk`, a community aimed at supporting the cultural sector online. It aims to provide the 'Latest news, exhibition reviews, links, event listings and education resources from thousands of UK museums, galleries, archives and libraries, all in one place'. Again this is not artist centric, this site's data is more event related but it does provide a number of Web feeds that allow its data to be automatically included in other websites.

The Europeana project goes a stage further in relation to aggregation of Fine Art Information (Haslhofer *et al.*, 2010). Europeana was launched in 2008, with the aim of 'making Europe's cultural and scientific heritage accessible to the public'. The portal `http://europeana.eu/portal` gives access to different types of content from various cultural institutions throughout Europe, and is funded by the European Commission. The information presented via this portal *is* in fact artist centric in that it 'makes it possible to bring together the works of a painter with, for example, relevant archival documents and books written about the artist's life'. It is also greater in extent and coverage given the large number of institutions which have taken part by allowing the inclusion of their digital content.

Each of these projects has been made possible as a result of the positive collaboration of the institutions, galleries and libraries involved. The information presented via each of these websites has been carefully selected and curated by the institutions that have 'opted-in' to the projects. With the exception of Culture24, it is not yet possible to programmatically retrieve the data available via each of these websites, for re-use. The data behind each of these websites is 'open' in the sense that the participating institutions have made it available for non-commercial use, but it is not freely available for programmatic consumption at large via an API or other form of Web service.

This is in contrast to the domain of Books and Music where data is made available in machine readable formats, thereby lending itself to the creation of near-comprehensive Web catalogues of information. `MusicBrainz.org`, for example, is a site that acts as 'a community music meta-database that attempts to create a comprehensive music information site', and which provides data about music to many other

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

websites and applications. The data which is aggregated on this site is then utilised by companies such as *The Guardian* and last.fm for their own music-related Web pages, and is augmented with additional and related data. This level of online data aggregation ultimately provides the Internet user with accurate, informative and high-quality information about music and sets a standard of online information provision that has become the expected norm in this domain.

A similar situation has arisen in relation to the domain of books, largely as a result of Amazon's book API. It is now possible, on many different websites, to view the complete catalogue of works by a particular author, related book reviews and other relevant data. Sites such as Librarything.com provide a near-comprehensive information service by combining open data from Amazon as well as hundreds of public libraries. Users can browse online the extensive book catalogues and can search by author's name, book genre, titles, subject-matter and even on the basis of 'most popular'.

It is clear that the online user experience in relation to browsing Fine Art Data cannot currently match that experienced within these two domains, given that the information remains largely distributed; and the reason for this is the very particular set of problems presented by the nine distinguishing features of Fine Art Information, as identified above.

2.1 Contribution

The aims of this project are based on the portion of the Web which is constructed from hyperlinked pages of visual information, text and images. Given that the original Web was for text-based document sharing, this is a considerable amount. These pages are either viewed online via HTML or word processor-based presentations, or can be downloaded in many formats. The retrieval of this Web information retrieval relies on the viewer being able to find it, usually through hyperlinked indexing systems. The challenge presented to information retrieval systems is to produce a reduced set of data from a larger collection to satisfy a user's information need. Some institutions have tackled this directly, as reported by Cahill (2009). This may result in excellent institute systems but it is unlikely to be a global or even portable solution.

There is a need for an indexing model based on the content users and purpose. Various virtual museum approaches have been addressed and implemented. Some are highly specific and based on a single institute as considered in Hertzum (1998) or institute groups as examined by Schweibenz (1998). Both these lead to individual efficiency and give valuable insight to the construction of a virtual museum. However, these approaches do not separate the institute from the data and therefore the portability of such models is limited to institutes with similar characteristics. An approach that is more compatible with the resources in question and more likely to remain viable with the ever growing and changing Web is one with a contextually broader view. It should be concerned with what it is indexing more than who holds the data. It must still examine the general needs of the relevant institutes, predict their future needs. All this must consider what is being indexed from a content perspective as discussed by Dyson and Moran (2000).

Our proposal consists of an API that requires an index of documents to be assembled using standard Web crawlers or by using available APIs. Knowledge of the structure of Web documents, which are reliant on HTML or related languages, allows their content to be automatically indexed. The API must also be able to review previous content owing to the changing nature of the data, as previously mentioned.

This paper reviews the available approaches, discusses the prototype Museum API 'ArtBridge' and presents the case for a generic Museum API. This gives the framework for interaction between independent Museum APIs that adhere to the generic one.

3 Motivation

The Web as a source of reliable information has already become unwieldy and at times unreliable; it is the sheer scale of this resource that presents the biggest challenges for individuals, businesses, organisations and developers alike. Visiting individual web pages to look for information is an inefficient use of time and energy, and although search engines can speed up the process, there is a growing need for the intelligent aggregation of topic-related information. It is Web technology that is not only driving and facilitating the increasing culture of open data, but also enabling us to make sense of it via applications that combine and

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

enrich data from different sources including websites, databases, Web/news feeds and spreadsheets, thereby creating new digital content in the form of 'mashups' (Merrill, 2006).

In essence, these applications transform raw data into understandable information by presenting it in a way that explains or visually maps out the story behind those facts or figures. Or as Rusbridger (2009) states, 'The web has given us easy access to billions of statistics on every matter. And with it are tools to visualise that information, mashing it up with different datasets to tell stories that could never have been told before'. The domain of Fine Art Information, however, appears to have been neglected.

Whilst it is relatively easy to find sites that aggregate *event* information related to exhibitions and what's currently on in the art world, it is less easy to locate information about a particular artist and his or her work. Information about Fine Art itself is widely distributed, and difficult to find without very specific targeted searching. For example, how do we begin to answer questions such as 'Where can I see artworks by the Scottish Artist Joan Eardley?'. A keyword search on Google for the name 'Joan Eardley' returns 293 000 results, including the following:

1. Aberdeen Art Gallery's online collection has 162 images of paintings and drawings by the artist;
2. Google Images contained 5740 results—only the first 10 pages contained relevant information;
3. the BBC portal features the town 'Catterline' on the programme 'Coast', and referred to Joan Eardley having painted there;
4. Wikipedia—information about the village of Catterline and it's 'notable inhabitants', which included Joan Eardley;
5. a Wikipedia biography of the artist;
6. Amazon.co.uk—a book about Joan Eardley by Cordelia Oliver;
7. *the Scotsman* newspaper published an article dated 2007 about the artist; and
8. *the Press and Journal*, a regional UK newspaper, published a newspaper article about the recent sale of an Eardley painting.

This exercise demonstrates that to find the information required about this artist, it is necessary to browse at least eight different websites. The information is there, and exists on the Web, but is spread throughout distributed sources. This problem becomes more pronounced as the Artist's notoriety increases: a keyword search in Google for the name 'Pablo Picasso' returns 21 200 000 results as of December 2011.

This raises the question as to how to bring all of this information together to make it accessible from one place. (Ayers & Watt, 2005: 4–5) have aptly summed up this situation as follows—'Most of us live in homes where water comes to us, rather than us having to travel to the water. It makes a lot of sense that information, too, should flow to us. It avoids the repetitive actions of going to visit individual Web sites and, if done well, achieves easier, more efficient and more effective access to information'.

At the moment, information about visual artists and their work is held in separate sources such as in the proprietary archives of public galleries or museums, or in private galleries. These sources are distributed, heterogeneous and often unrelated (in the sense of not being linked together) (Baca, 2002). Some galleries only represent a handful of artists, whilst some artists exhibit their artworks at numerous galleries, spread throughout the world. There is no pooling of resources to provide a more comprehensive presentation of information in relation to:

1. the bibliography of the artist;
2. the images of the artists' artworks;
3. the exhibitions in which the artist has taken part (and the artworks included in those exhibitions);
4. news articles of relevance to, or about, the artist or their exhibitions;
5. reviews of artists and exhibitions; and
6. information about the places that relate to the artist or the artwork.

Fine Art Information is not 'artist centric' when taken as a whole. Although it is possible to search specific collections for artworks by a particular artist, it is not possible to find any one source that lists all of the artworks associated with a particular artist's name, and where it currently resides. This is in contrast to

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

the domain of books, for example, where numerous online catalogues provide the means to search by author to find a full list of their works, related book reviews and even book-cover images. Librarything.com is one such catalogue—this not only brings together information from hundreds of distinct library catalogues, but also creates an online community of book lovers.

The principal distinguishing feature between this domain and that of Fine Art is the ISBN number. This enables books to be uniquely identified and catalogued, and at the same time facilitates aggregation of related information. Whilst museums and galleries do catalogue artworks according to certain minimum standards, there is no ubiquitous standardised method of identifying a particular artwork that would be equivalent to the ISBN number (Baca, 2008). There are accordingly huge variations from gallery to gallery in terms of the quality and completeness of the data recorded. There are also problems associated with both differing formats of data and different languages. Whilst collaborative attempts have been made to specify a schema to support data interchange between public galleries, such as that devised by the Getty Institute[1], but these schema are not universally adopted.

Overall, Fine Art Information can be characterised by a number of distinguishing features, which, taken together, present a unique set of challenges in so far as ingathering and organising that information using traditional retrieval techniques is concerned. The information is characterised by:

1. its dispersed and distributed nature;
2. a huge variation in *quality* and *quantity*, depending on the artist's notoriety and the source of the information;
3. transience, particularly in relation to contemporary art and living artists;
4. heterogeneous formats;
5. the restrictive nature of repositories or archives (i.e. the information is not freely available for re-use);
6. a lack of uniformity in relation to the classification of images and art terms;
7. the fact that words (usually the artist's name and subject) are used in a non-standardised way to identify and search for images (Baca, 2002; Manning *et al.*, 2009: 178);
8. incomplete or inconsistent data; and
9. restrictions related to copyright and re-distribution of images of artworks.

In view of these difficulties, Art Information remains heterogeneous, distributed and difficult to aggregate except in relation to event information: for example, New York Art Beat (NYArtBeat.com) is a site devoted to listing all art and design events in New York, and claims to be a 'Smart data organisation with events sorted by media, schedules, and location, as well as event lists like Closing soon, Most popular, Open late, and Free'. It aggregates relevant art reviews and operates an intelligent tagging system that permits users to search easily for events of interest to them. In the true spirit of open data, it also provides an API[2] that permits the inclusion of its information in other websites. Given the above set of problems inherent in Fine Art Information, it is not possible as matters currently stand to aggregate it with the same ease with which event information is brought together.

The question arises, then, as to how the problems of heterogeneity and distribution might be overcome in this domain, withstanding the inherent difficulties.

It would not be desirable or practical to seek to create a comprehensive centralised database of Fine Art Information by its very nature, Fine Art Information is constantly changing. Every day artworks are created, purchased, sold, loaned and even discovered; news stories or reviews of exhibitions are constantly being published; Fine Art Information is not a static data set and as such does not lend itself to a permanent amalgamation. It takes time to put together a data set from a large public collection and often by the time that data set is established, it is already out of date. To carry out this task manually for each of the public galleries in the world would be a task without end. There are also issues related to copyright that would prevent the centralised storing of images without express permission from each copyright owner.

[1] Available at `http://www.getty.edu/research/publications/electronic-publications/cdwa/cdwalite.html`

[2] Available at `http://www.nyartbeat.com/resources/doc/api`

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

3.1 From data to information via Web technology

It has been said that good data visualisation starts with asking questions about what story the data can tell, and what is interesting about it (Fry, 2007: 4). If it can communicate the story told by the relative data, then the visualisation is considered to be a success.

In the present case the aim would be to aggregate Fine Art Information in a more 'artist-centric' way, and therefore the question that needs to be answered is 'where can I see artworks by' a particular artist. This is the story that needs to be told by Fine Art Data if it is to be given the same online treatment as the domains of books and music. To demonstrate how Fine Art Data retrieved from museum APIs might be utilised, it is useful to examine the way in which other websites have aggregated open data in other domains, to tell a particular story.

It is clear that visualisation can be much more than just a graphic representation of a static set of facts and figures—when one data set is combined with and enriched by another, with this newly combined information then being translated into a *dynamic* graphical format, the resulting application becomes a powerful communication tool capable of providing instant answers to the user's specific query.

The best way to highlight this is by way of example: the website `mysociety.org` is run by a charitable organisation that seeks to build websites to promote openness and democracy in public life. One of their most popular projects is TheyWorkForYou.com—this website takes the current list of Members of the British Parliament and maps that data to the UK postcode data set, allowing users to enter their own postcode to search for details of their parliamentary representatives. From there, users can see whether their MP has been present at a particular parliamentary debate and view details of what that MP has actually said in debates, this information being derived directly from Hansard, the official archive of daily Parliamentary debates. Users of the site can even choose to be alerted by e-mail whenever a particular MP speaks in Parliament and can e-mail the MP directly. The website also combines two further data sets—the Register of Members' Interests and Expenses data—both of which can be searched at the click of a button. In bringing all of this public information together in a simple user-friendly interface, this website makes it easy for constituents to keep tabs on their Parliamentary representatives and, in doing so, increases their representatives' accountability.

This is a good example of Web technology being used to bring together publicly available data in a way that transforms it into a consumable story, and presents it in a more user-friendly way. As David Whiteland from `mysociety.org` said very succinctly during the writer's discussions with him (in London in July 2010) technology changes 'data to information'[3]. Data in isolation is just data, but data linked to other related data that is presented in a user-friendly way becomes useful information or knowledge. When that knowledge is made available via the Web, its potential audience is almost global.

There are many different ways to present information via a Web interface, and a review of relevant websites suggests that there are four different levels of dynamism of data applications that range from simple graphical illustrations, to fully interactive websites. The four levels are as follows.

First, there is the simple unchanging graphical representation of a static set of facts—the subway map being a good example of this. That simple image presents a very refined view of geographical information, station locations and routes in a way that makes it easy to plan a journey. The data set upon which this visualisation is based is relatively static over time (unless of course a rail line is extended or a new station built) as is the visualisation itself. An example of a Web-based visualisation at this level might be the Linked Open Data Cloud that is an interactive visualisation of all the linked data sets that exist at a certain point in time. This can be accessed online (at `lod-cloud.net`) and when clicking on any of the circles containing the name of a data set, the user is taken directly to the source of that data set. A simple yet highly effective visualisation of a large data set.

The second level is a more dynamic graphical representation of a static set of facts. An example of this can be found on the website `wheredoesmymoneygo.org`. This website seeks to 'promote transparency and citizen engagement through the analysis and visualisation of information about UK public spending'. Via its 'Dashboard'[4] application (see Figure 4), this website provides a stylish visual record, built using

[3] In conversation in London, July 2010 at *The Guardian* offices, Kings Place, London.

[4] The Dashboard can be found at `www.wheredoesmymoneygo.org/bubbletree-map.html#/~/grand-total--2010-`

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

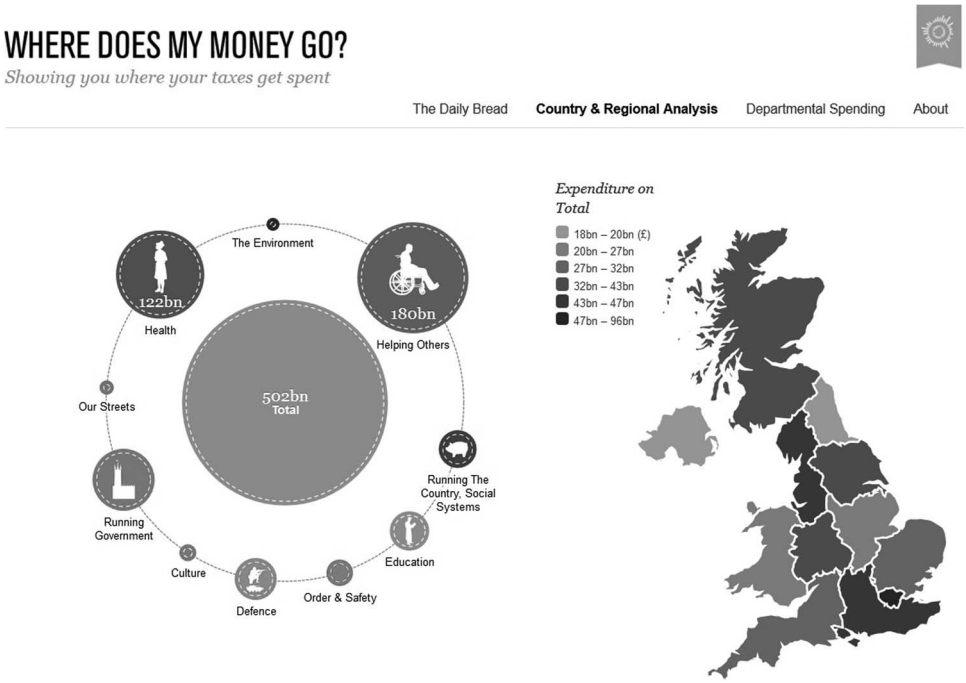

Figure 4 The 'where does my money go' dashboard (screenshot from `www.wheredoesmymoneygo.org`)

Flash, of where public money has been spent, by year. The data upon which the visualisation is based is historic in nature (i.e. unchanging) and ordinarily this set of financial figures would make for very dull reading to the majority of people. However, here we see the power of dynamic interactive graphics—the apparent simplicity of the coloured circles communicates a complex set of statistics that is capable of being understood by non-statisticians. It provides a timeline allowing for an easy visual comparison of spending by year, from 2003 to date, and allows users to click on each area of expenditure (such as Health) for a further breakdown of the spend involved that year.

This is a very good example of technology and design working together to communicate the story being told by an otherwise overwhelming set of public data.

The third level of visualisation is similar to the second in that it involves a relatively static data set, but with the addition of a further layer of interactivity that allows the user to modify the view of the information by entering search criteria, or choosing a refinement of the data from a list. An example of this would be the website *nukeometer.com* built by Adam Charnock that uses JQuery, Google Maps API and a data set comprising the locations of all nuclear warheads in the world, derived from a news article in *The Guardian* (Rogers, 2009). It provides a very simple interface into which the user enters their current city and country. From there the application lists the number and location of the warheads within range of that city on a deceptively simple screen which displays quite a shocking message.

The fourth level of visualisation involves two distinguishing features—a non-static data set, that is, a set of facts or information that is constantly changing and which requires regular dynamic updates; and second, the participation of and contribution by the user (see Figure 3). Such applications are complex, often involving many different features and/or data sources, and even the creation of a community of interests or social network. They invariably involve writing scripts to dynamically update the appropriate web page, either directly from Web resources or from a database into which the information has previously been stored. This type of application or service has two main features—the ever-changing data set and user participation.

One website which demonstrates such dynamism is LibraryThing.com, previously referred to. This is, in the writer's view, a site that exemplifies what can be done using intelligent Web technologies—it brings

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

together publicly available data about books whilst creating a community of people sharing a common interest in reading. At the same time it facilitates the creation of a comprehensive catalogue of book information that is constantly being augmented by the input of millions of users.

This site makes use of data from Amazon's book API, plus book catalogues from the Library of Congress in the United States and, according to the website, '690 other world libraries'[5]. In essence, this website is both a cataloguing facility and a social network for book lovers, requiring users to log in to their own profile page. It makes use of the constantly updating data stored by each of the source libraries, but enriches that data by linking them to other books, reviews, users, tags and images, thereby creating an enhanced database of books in addition to a social network of readers. It also provides recommendation services based upon other users' reviews and even their conversations, in addition to a 'zeitgeist' recommendation feature that collates site statistics, using them as the basis for dynamic listings such as 'most read', 'most reviewed' and 'top author'. With over one million users adding content to the site, and around 50 million books in its system, the data upon which LibraryThing is based, is constantly being augmented and updated.

This website demonstrates how open data can be used in a very creative and dynamic way by bringing together disparate library catalogues of books, transforming them into a more comprehensive and updating repository of book and book-*related* information with a vibrant community of users. It has recently been argued that while catalogue records will continue to be the kernel of bibliography, thoughtful reviews will assume greater importance (Wagner & Weibel, 2005).

Thus, applications that combine both data and social interaction can serve not only to present that data in a consumable form, but they can become platforms through which the data can be enriched by user interaction, thereby 'acting as collective intelligence gatherers' (Bell, 2009: 5). Such a situation is only made possible by Web technologies that permit both the dynamic aggregation and dynamic searching of vast and increasing amounts of open data. `LibraryThing.com` has access to a large array of online and trusted data sources relating to literature of all kinds, which enables it to provide a near-comprehensive online cataloguing facility in relation to books. If a similar volume of programmable Fine Art Data was made available online, there would be no barrier to providing a similar aggregation and cataloguing facility in relation to the works of art created by the artists of the world. This would ultimately improve the experience of those searching for Fine Art Information on the Web whilst at the same time would increase the accessibility of many museums' online collections.

4 A possible solution: the Museum API

It is the writer's view that in order for the aggregation of Web-based Fine Art Information to be feasible, it is necessary to have a greater degree of machine capabilities of Fine Art Data. Most public and private art galleries have a publicly accessible website displaying digitised images of their artworks. The data is already in a format that would lend itself to inclusion in the world of open data. However, the means to automate the accurate processing and aggregation of this information is lacking. As previously indicated, the creation of a centralised repository of all Fine Art Information is neither feasible nor desirable given the fluidity of the data itself. A more pragmatic and dynamic solution is required, and it is the writer's view that this could be provided by the creation of APIs to each digital repository of Fine Art Information, whether it is a public collection of art objects or a private commercial gallery.

The provision of an API to a museum collection, or indeed to a commercial gallery, has a number of useful features for both the museum and the developer alike, which can be summarised as follows.

First, the terms of use of the data are made clear from the outset by the institution providing the Web service. The images of artworks in respect of which public distribution is prohibited, are generally excluded from the API, or included with very low-resolution files. Further, there are limits on the purposes for which the data can be utilised, limits on the number of calls to the API that can be made per day, and terms requiring that specific permission be obtained in certain circumstances, for example, where the institution's logo is to be used in an application.

[5] `http://www.librarything.com/`

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

Second, once a museum has carried out the work to set up the API, it need not spend any further time dealing with requests for, or putting together, specific datasets given that the information can then be obtained programmatically (and in the desired format) using appropriately constructed URLs. A very efficient use of resources all round, and a pragmatic means to overcome the hurdles posed by the lack of homogeneous formats.

Third, as far as the developer is concerned, using carefully constructed URLs it is possible to retrieve only the data that is required, in the knowledge that as it comes from a trusted source, it is likely to be highly relevant and of reliable quality. There is the added advantage that where a Web service utilises persistent URLs, the resource at that specific address will always be up to date so long as the API is properly managed.

And finally, for relatively minimal effort, the museum can automatically expose its collection to a greater audience on the Web by making its data available to enthusiastic Web developers as well as museography specialists. The result will invariably by a plethora of interesting and novel Web applications, each of which publicises (for no effort or cost to the institution) the content of the museum's collection. It is only necessary to look at websites such as programmableweb.com to see this in practice—as at July 2011, this site indicated that there were 2243 Web applications utilising the Google Map API (available at `code.google.com/apis/maps/index.html`). These figures speak for themselves.

5 Implementation: ArtBridge

A simple example of what might be achieved in the domain of Fine Art is provided by the Web application ArtBridge (see Figure 6), a project of the Robert Gordon University in Aberdeen (available at `www.comp.rgu.ac.uk/ArtBridge`).

The system has been deployed using the MAMP 1.9 (Mac, Apache, MySQL, PHP `http://www.mamp.info/`) software stack under the Mac OS X 10.5 operating system.

Our system has abstracted the connections to various Web resources to retrieve the results related to that search term, decoupling the logic of the query from the real infrastructure dependencies. The system processes the information, organising it into an artist-centric file containing relevant URLs that point to Web resources relevant to that artist. When a name is input via the Web interface, the system retrieves the file of relevant URLs and obtains the resources from those links, re-presenting them on the web page for that particular artist. In this way, the system aggregates relevant information in relation to artists and provides access to that information from each of the distinct sources, in one place.

This application, written in Java, retrieves information from mainstream Fine Art sources:

1. New York's Brooklyn Museum API;
2. London's Victoria & Albert Museum API;
3. *The Guardian*'s Open-Platform; and
4. Aberdeen Art Gallery & Museum.

It analyses the relevance of the data retrieved from each resource, and stores the URLs for the relevant resources in a separate XML file for each particular artist.

The ArtBridge system allows the user to choose the name of a particular artist, and from there a Web application, written in ActionScript 3.0, then displays the relevant data retrieved from each of these four Web sources using the previously stored URLs. Each item of data is stored in a separate box within the display, and includes images from each of the Museums' digital repositories, and relevant news articles from *The Guardian*'s Web service. All of these resources are retrieved dynamically from the URLs and therefore display the current information available at that particular URL.

It is, however, designed in such a way that additional sources and formats of information can be incorporated in the system, without difficulty.

The main application can be run from a simple GUI. Each option is managed by the StartApp class, which contains the main method. This class retrieves the input search terms from either the GUI or command line, and creates a new StartSearch object which is the kernel applicative object through which

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

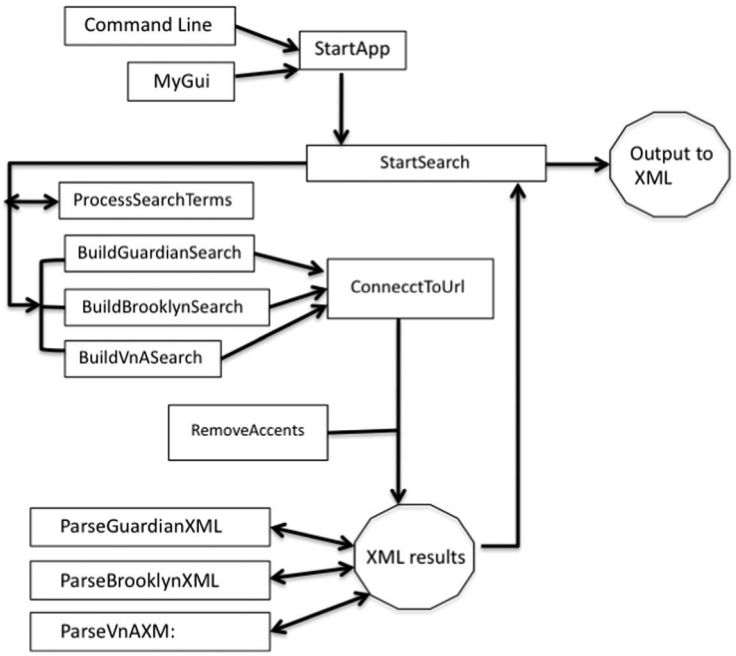

Figure 5 Simple block diagram of the ArtBridge system

all of the information in the system flows. It creates the 'BuildXXXSearch' objects which manage the construction of the URLs for each API, and manages the XML output. This is illustrated in Figure 5.

The system is designed so that a 'BuildXXXSearch' object is instantiated for each API (*Guardian*, Brooklyn, Victoria & Albert Museum, etc.), with its own class design being dependent upon the idiosyncrasies of each of the repository APIs but with common attributes being inherited. In this way the details of the construction of specific URLs for each API are encapsulated within separate classes, and it also means that the system can be easily extended to cope with additional APIs simply by creating a new 'BuildXXXSearch' derived class for that API, for example: 'BuildGuardianSearch(String key, String artistName)'.

Where key represents the authentication parametric values for a given API (in the above example, *The Guardian*'s) and artistName is a string which is then parsed and converted into a searchable artist entry (typically formed as a first–last name pairing). A correct parsing is crucial because otherwise a seemingly straightforward query on 'Vincent Van Gogh' could include details of all artists with a first or middle name of 'Vincent' or 'Van' as well as personalities with 'Gogh' as a surname (e.g. Theo Van Gogh). Furthermore, open non-qualified search terms can often return a number of seemingly random artist's names (e.g. 'La Oreja de Van Gogh', a Spanish pop band).

When the search term has been processed in this way, the BuildXXXSearch object then puts together a specific URL String to enable the system to query the API. It is noted that the relatively large number of string variables reflects the complexity of queries that can be made to the API of the different information repositories.

Query refinement is API dependent and therefore the BuildXXXSearch objects must contemplate the subtleties of the repositories. For example, for *The Guardian*'s API, we could change the newspaper section in which the data might appear. In the present case we are interested in '`artanddesign`' or '`culture`', but there are over 50 different sections that could be queried. A typical query URL looks like:

```
content.guardianapis.com/search? q=pablo+picasso&section=artanddesign&
format=xml
```

Located immediately after the question mark, the tags part of the URL corresponds to *The Guardian*'s classification of news content. There are literally thousands of different tags by which queries can be

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

refined. In the present case, the hard-coded query is for articles, reviews or news but not obituaries, but this could be changed to look for only reviews in a particular section, for example. The URL String is therefore assembled and from that, a new URL object is created. The public method getUrl() of the BuildGuardianSearch class returns the URL object in question to the calling StartSearch object. The BuildBrooklynSearch and BuildVnASearch classes operate in a similar way. The information returned by each API is then parsed, with the relevant segments of data being stored in custom data objects. What the data objects will be storing, in effect, is a list of URLs that point to resources relevant to the particular artist. Some of those URLs will be known in advance, such as the links to biographies and links to galleries with which the artist has an association. Others will be retrieved from searches to the relative APIs. The object is not to acquire and store the actual images or text in a centralised database, but rather the URLs that point to these resources. The reason for this is threefold.

First, there are complex copyright issues attached to the storing of images of artwork; second, it is not usually permitted (in the applicable terms of use) to store or cache data retrieved via an API for longer than 24 hours; and third, given that information on the Web is capable of changing rapidly, it is necessary to ensure that any data displayed via the application is up to date and relevant. It would therefore be desirable, in view of these constraints, to access the resources dynamically as required to ensure that the information displayed is always up to date. Finally, after the information is retrieved from the given repository the information is then rendered as a Web interface that provides a degree of interactivity and manipulation of the views of the data in question. The user is able to choose an artist's name from a list and view the relevant information. Consideration required to be given as to how that interface was to be built, what scripting languages were to be used, and how those were to be deployed and tested. Ultimately, the chosen method has to use XSLT/CSS stylesheets, applied to the appropriate XML document selected using PHP as the scripting language.

6 Evaluation

It is important in this particular project to work with real data from the relevant APIs in order to obtain accurate feedback from the content of the results returned. Open data from a select group of artists has been used to test the operation of the system and indeed in the final evaluation of the overall system. The artists are Pablo Picasso, Henri Matisse, Tracey Emin, Vincent Van Gogh, Jackson Pollock, Claude Monet, David Hockney and Andy Warhol, all of whom have data in at least one of the APIs used in the system. To evaluate the actual functioning of the system, each artist's name has been input via the system GUI, and automatic searches were then carried out consecutively in the three APIs (*The Guardian*, Brooklyn Museum and Victoria & Albert Museum). For each artist, the results were output to the console so that an immediate assessment of the results could be made.

As summarised in Table 1, the number of URLs are highly focused and relevant reporting actual references to the artwork and life of a given fine artist, as opposed to an assorted collection of loosely related pages. As an illustration, a simple Google search on Vincent Van Gogh on *The Guardian* site (`guardian.co.uk`) produces over 3000 results in stark contrast to the seven URLs reported.

From the main display of information illustrated in Figure 8, the user can instantly see which of these three museums hold artworks by the particular artist chosen, and can view images of them if available. It is also possible to read relevant art reviews or news articles of relevance to that artist. A screenshot of the data relating to Henri Matisse as presented by ArtBridge user interface appears in Figure 8.

Each image block, when hovered over by the mouse, displays the title of the artwork, and the city and name of the museum to which it belongs as shown in Figure 9.

Clicking on the image displays either the full-sized image or the full news article. The information being brought together from these four distributed resources is re-organised in an artist-centric way, and avoids the need on the part of the user to separately visit these four different websites. However, given that the application is making calls to the relative APIs each time that an artist's name is selected, it is increasing Web traffic to that Web service, whilst at the same time increasing public access to the content of each institution's online digital collection. It also demonstrates a possible answer to such questions as

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

Table 1 System evaluation results using a set of eight Fine Artists with *The Guardian*, Brooklyn and Victoria & Albert Museum (V&A) Application Programming Interface

	The Guardian	Brooklyn	V&A
Pablo Picasso	25	10	3
Henri Matisse	8	0	4
Tracey Emin	11	1	1
Vincent Van Gogh	7	2	0
Jackson Pollock	1	0	0
Claude Monet	4	5	0
David Hockney	8	0	11
Andy Warhol	11	10	2

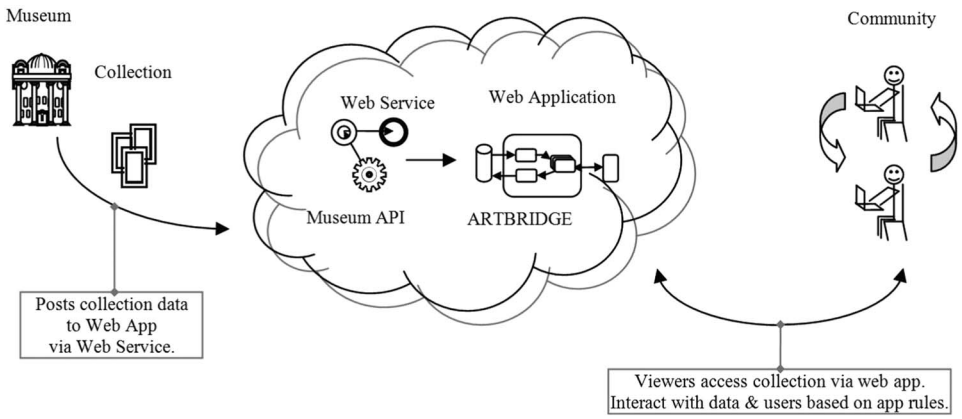

Figure 6 A representation of ArtBridge and the Museum API (Application Programming Interface)

'where can I see artworks by Henri Matisse?' and in doing so highlights the potential benefits of increasing the aggregation of Fine Art Data.

6.1 Discussion

At the moment, the scope of the application is limited by the number of online data sources available, although it is designed so as to be capable of the modular addition of many more. As the number of data sources increases, so too will the quality, accessibility and usability of the Fine Art Information provided online (see Figure 7).

If a project such as Europeana was to provide an API allowing for online access to its million-plus digitised items (and this is mentioned as a possibility on their website), the effectiveness and utility of applications such as ArtBridge would increase exponentially. Further, if every public museum or gallery in each of our major cities were to allow access to their online data via an API then it would be possible to imagine a situation where the Web of Fine Art Data is aggregated to such an extent that a near-comprehensive catalogue of many artist's works could be viewed, reviewed and augmented at a single location rather than the thousands of websites over which it is currently distributed.

These are big, but not inconceivable, 'ifs', which if realised would bring the quality of online information in this domain up to the standard currently enjoyed in the book and music domains. The Web technology needed to effect this transformation already exists but is under-utilised in this area of Fine Art Data.

It is the writer's view that the museum API can act as a knowledge bridge between the distributed online digital repositories of Fine Art Data by providing for programmatic access to that data, and in doing so it

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

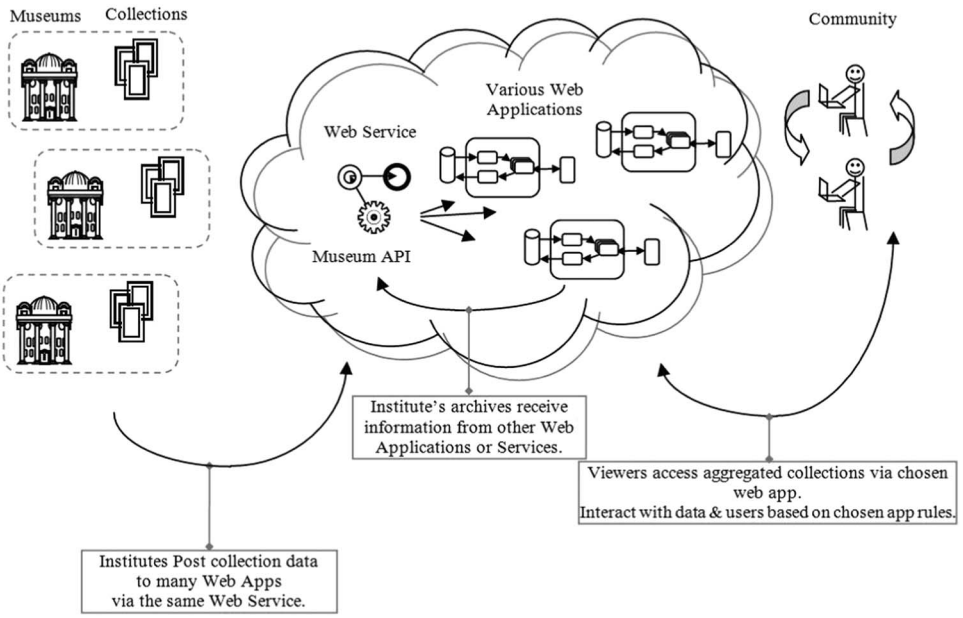

Figure 7 Generic museum standardisation for interactive APIs (Application Programming Interface)

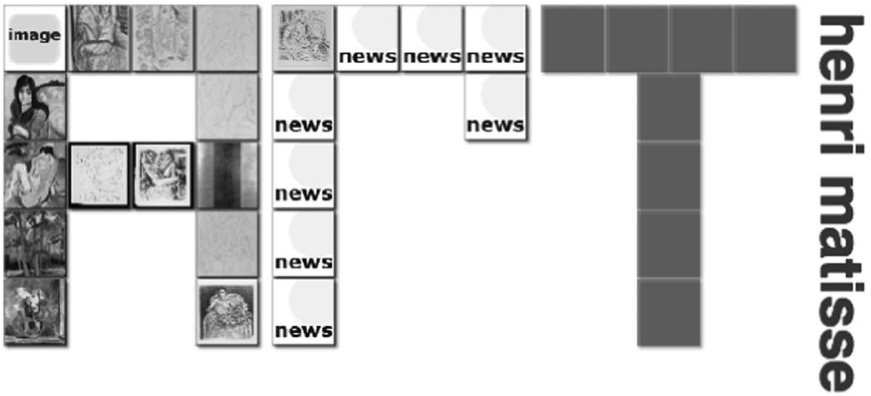

Figure 8 ArtBridge user interface—Henri Matisse example

Figure 9 ArtBridge image block

can facilitate its aggregation and contribute to the improvement of the online experience in this domain. It provides a pragmatic solution to the problems inherent in this domain, whilst at the same time increasing the online accessibility of each of the sources of digital Fine Art Information.

7 Conclusions

In this paper we have been concerned with the specific context of highly independent, heterogeneous and distributed sources of Web information on Fine Art and Fine Artists. The problems addressed were those

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

of large-scale information indexing and retrieval. The domain data occurs in various types; long term/static such as biographies and artworks, transient such as art collections and time specific such as exhibitions. The independence of the data naturally results in a heterogeneous organisation and structure so that separate data repositories cannot freely interact. This often creates duplication and contradiction in data. Contradiction is especially true when time dependence affects the relevance of the data.

We established that the data is not the problem. The problem relates to different methods for information retrieval which result in specific, and non-interactive, APIs being created for the same type of data. Problem examples are shown through the need to source and filter data independently from multiple sites to obtain the required information.

The paper established the continuity of data organisation in other domains, relating to pre-Web organisation such as library systems. It highlighted distinguishing features such as ISBN and showed the contrast in universally accepted techniques within Fine Art cataloguing. In contrast to long-established methods, the rise Web participation in numbers and interaction is shown to be fundamental to data retrieval.

We presented a proposal for the development of a solution in terms of the Museum API. This examined the Fine Art domain, presented the argument for a generic API and offered a prototype 'ArtBridge' that supports the aim of the Dublin Core to enable mapping of disparate data sources. This is supported by a review of different Web applications, which have been created using open data sources.

A possible solution, The Museum API, is examined in terms of the current state of Fine Art Information, highlighting the problems. Support for our proposal ideas is currently observable through a small number of specialist, collaborative, aggregation projects.

From the previous work, the requirements of the Museum API are established. Having already established that the data is not a problem, this shows that the data format is also not a problem; the means to aggregate is a central problem and a solution relies on automating an accurate process. The proposal is against a centralisation approach, which would be an impossible, never-ending task and would also require a top-down authority throughout the Fine Art domain. Instead, it supports a pragmatic solution of an API that encourages community participation and open source development. This approach allows the independent creation of APIs that can interact with each other.

We obtained four solution steps from the requirements and these are; clarity of data and data use, API robustness (maintenance free), establishing trust in sources to validate the URL feeds, and automation of process to minimise the individual effort required for individual institutes to participate.

Finally, the power of data visualisation, which will be obtained via such an API, is presented with examples in several different domains. This leads directly to our prototype solution, where we present ArtBridge as a Web app built using several Museum APIs. This counters the problems we have observed and fits the requirements of the solution. Future work will be establishing a full system that completes all four solution steps.

Acknowledgements

This research was partly supported by a Scott Trust Technology Bursary. The authors acknowledge use of the services and facilities of the School of Computing at the Robert Gordon University.

References

Artz, D. & Gil, Y. 2007. A survey of trust in computer science and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web* **5**(2), 58–71.

Ayers, D. & Watt, A. 2005. *Beginning RSS and Atom Programming*. ISBN: 978-0-7645-7916-5. Wiley.

Baca, M. (ed.) 2002. *Introduction to Art Image Access: Tools, Standards, and Strategies*. ISBN: 0892366664. Getty Research Institute.

Baca, M. (ed.) 2008. *Introduction to Metadata*, 2nd edition. ISBN: 0892368969. Getty Research Institute.

Baeza-Yates, R. 2003. Information retrieval in the web: beyond current search engines. *International Journal of Approximate Reasoning* **34**(2–3), 97–104.

Bell, G. 2009. *Building Social Web Applications*. ISBN: 0596518757. O'Reilly Media.

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press

Buchanan, F., Capanni, N. & González-Vélez, H. 2011. Fine artists of the world unite: bridging heterogeneous distributed open data sources of fine art. In *i-Society 2011*. IEEE, 224–229.

Cahill, K. 2009. Building a virtual branch at Vancouver Public Library using Web 2.0 tools. *Program: Electronic Library and Information Systems* **43**, 140–155.

Dyson, M. C. & Moran, K. 2000. Informing the design of web interfaces to museum collections. *Museum Management and Curatorship* **18**, 391–406.

Fry, B. (ed.) 2007. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. ISBN: 978-0596514556. O'Reilly Media.

Haslhofer, B., Momeni, E., Gay, M. & Simon, R. 2010. Augmenting Europeana content with linked data resources. In *I-SEMANTICS'10*. ACM, 40:1–40:3.

Henzinger, M. 2001. Hyperlink analysis for the web. *IEEE Internet Computing* **5**(1), 45–50.

Hertzum, M. 1998. A review of museum web sites: in search of user-centred design. *Archives and Museum Informatics* **12**, 127–138.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C. & Lee, R. 2009. Media meets semantic web: how the BBC uses DBpedia and linked data to make connections. In *The Semantic Web: Research and Applications*. LNCS **5554**, 723–737. Springer Verlag.

Manning, C. D., Raghavan, P. & Schtze, H. (eds) 2009. *Introduction to Information Retrieval*. ISBN: 0521865719. Cambridge University Press.

Mayfield, J. 2002. Ontologies and text retrieval. *The Knowledge Engineering Review* **17**(1), 71–75.

Meng, W., Yu, C. & Liu, K.-L. 2002. Building efficient and effective metasearch engines. *ACM Computing Surveys* **34**(1), 48–89.

Merrill, D. 2006. Mashups: the new breed of web app. *IBM Web Architecture Technical Library*, 1–13.

Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V. & Varga, L. 2008. The provenance of electronic data. *Communications of the ACM* **51**(4), 52–58.

O'Reilly, T. 2007. What is Web 2.0: design patterns and business models for the next generation of software. *Communications & Strategies* **1**, 1–17. `http://ssrn.com/abstract=1008839`.

Proskine, E. A. 2006. Google's technicolor dreamcoat: a copyright analysis of the Google Book search library project. *Berkeley Technology Law Journal* **21**(1), 213–240.

Rogers, S. 2009. The world in active nuclear weapons, *The Guardian*, 6 July. `http://www.guardian.co.uk/news/datablog/2009/apr/06/north-korea-nuclear-weapons`. Accessed 21 May 2012.

Rusbridger, A. 2009. Free the facts: the Guardian's editor-in-chief on why open data matters, *The Guardian*, 10 March 2009. `www.guardian.co.uk/news/datablog/2009/mar/10/1`. Accessed 23 April 2012.

Schweibenz, W. 1998. The 'virtual museum': new perspectives for museums to present objects and information using the internet as a knowledge base and communication system. In *ISI*. ISBN: 3-87940-653-7. *Schriften zur Informationswissenschaft* **34**, 185–200. Hochschulverband für Informationswissenschaft.

Shirky, C. 2010. *Cognitive Surplus: Creativity and Generosity in a Connected Age*. ISBN: 1846142172. Penguin Books.

Snásel, V., Abraham, A., Owais, S., Platos, J. & Krömer, P. 2009. Optimizing information retrieval using evolutionary algorithms and fuzzy inference system. In *Foundations of Computational Intelligence*. Studies in Computational Intelligence **204**, 299–324. Springer Verlag.

Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E. & Giordanino, M. 2007. The usability of semantic search tools: a review. *The Knowledge Engineering Review* **22**(4), 361–377.

Wagner, H. & Weibel, S. 2005. The Dublin Core Metadata Registry: requirements, implementation, and experience. *Journal of Digital Information* **6**(2), 1–20.

Weibel, S. 1997. The Dublin Core: a simple content description model for electronic resources. *Bulletin of the American Society for Information Science and Technology* **24**(1), 9–11.

https://doi.org/10.1017/S0269888914000319 Published online by Cambridge University Press