Extending Domain-Specific Resources to Enable Semantic Access to Cultural Heritage Data

Paul Clough
University of Sheffield, UK
p.d.clough@sheffield.ac.uk

Neil Ireson
University of Sheffield, UK
n.ireson@dcs.shef.ac.uk

Jennifer Marlow
Carnegie Mellon University, USA
jmarlow@cs.cmu.edu

Abstract

Cultural heritage material often contains rich semantic information, which can be utilised for alternative forms of information access beyond keyword searching and browsing by subject categories. In order to provide such functionality it is desirable to annotate all the material in a collection with named entities and their relationships so that all the collection is available for semantic search. In this paper, we examine issues involved with automatic semantic annotation of information about artists from Tate Online using a pre-existing domain-specific structured resource (ULAN). In particular, we focus on extending ULAN's coverage of artists and their associated semantic properties (e.g. birth/death date, birth/death location) by applying focused crawling and automatic information extraction techniques to exploit semi-structured sources of information. This enables the cross-referencing of collections against a range of information sources, thereby improving visibility and end-user information access.

1. Introduction

A principal concern of many cultural heritage organisations is to provide access to their collections, both physically and, more recently, digitally. Similar to accessing digital libraries, many institutions provide users with both free-text searching of collection content, together with browsable semantic categories, such as subject type, also useful in organising items.Benjamins et al. (2004) highlight the value of semantics in the humanities domain, stating that most information-seeking in this area involves *"events, persons, and movements in a historical or cultural context".* Similarly, Hyvönen (2007) asserts that the cultural heritage domain is well suited to the creation of semantic portals. These can, among other things, (1) give an aggregated, global overview of heterogeneous content and (2) provide a more "intelligent" way of examining content through semantic linkages.

Advances in technologies, such as the Semantic Web (Fluit et al., 2005), have made it increasingly possible to search and browse for items in cultural heritage collections using richer sets of parameters based on *semantic* information. AsMaedche and Staab (2002) explain, machine-readable descriptions (as provided by Semantic Web technologies) can be utilised to facilitate finding, integrating and connecting information in a way above and beyond that which can be done with a simple keyword search. Hildebrand et al. (2007) outline the various elements of the semantic search process, which include construction of the query, execution of the search algorithm, and presentation of results. With regards to interface design matters, they mention both typical and more experimental visualisation techniques ranging from ranked lists, clustered result displays, tag clouds, cluster maps, and data-specific designs such as timelines.

To utilise Semantic Web technologies, semantic information must be made explicit and referenced to an underlying semantic model (i.e. an ontology). This is part of a process called semantic annotation (see Section 4). There are three general factors which require consideration when developing such a semantic model:

1. *User requirements:* does the model provide the necessary concepts to satisfy the information access needs of the user?
2. *Data modelling:* is the model expressive enough to capture the nuances of the data?
3. *Interoperability:* does the model provide the ability to interact (input/output information) with other models?

To illustrate these factors, consider a simple case where the collection contains information about artists, such as Joseph Mallord William Turner born on 23 April 1775. For some users and uses of the collection this information may be sufficient, however Joseph Turner may also be referred to by other variants of his name: J.M.W. Turner and Joseph Turner, the latter also being the name of a contemporary artist who worked in a similar style. It may also be desirable to know that he was a painter (in watercolour) and a printmaker. In addition, there is some dispute about his exact birth date and users may wish to know not only the alternative dates, but also the sources of that information. Rather than providing this information simply as text, and relying on keyword searches, if it is available for machine processing then it enhances the user's ability to access the collection by using relational (e.g. born before 1900) or Boolean operators (e.g. NOT a painter) on the information facets. To fully represent the data in this example the semantic model should be capable of expressing multiple potential values with a preferred value, certainty measures and provenance.

One of the guiding principles behind the development of the Semantic Web is interoperability: facilitating the interconnection between information published by different individuals and organisations. By providing the possibility to link all the data relating to a given instance of a concept (such as the artist J.M.W. Turner), synergistic benefits can be derived and the users can access a collective understanding of information available. The process of linking data in the Semantic Web involves allocating an identifier (generally referred to as a Uniform Resource Identifier or URI) to each resource (instance of the concepts) in a collection and then explicitly linking the resources, both within the collection and to resources in other collections. This is done using the Resource Description Framework (RDF) which, in its basic form, uses expressions, known as triples, to define the relationship (predicate) between two resources (subject and object). In recent implementations, triples have been replaced by quad expressions, with the fourth element being used to represent the context or source of the expression, which provides a means to represent provenance and trust within RDF (Carroll et al., 2005).

One organisation to embrace this principle is the BBC (Raimond et al., 2008, Kobilarov et al., 2009), they have developed a model of the concepts (an ontology) relating to their programmes (see BBC Ontology), and are publishing all their TV and Radio information with their associated URIs. This information is then linked into the Semantic Web and enables queries such as, "which programmes play music by artists born in 1950s London?", even though this information is not available within any single organisation (including the BBC).

In order to provide reliable semantic search, it is necessary to annotate all objects in the collection, otherwise those objects without associated semantic concepts will be not be accessible to users. Despite the growth in the amount of linked data available (from around 500 million triples in May 2007 to over 6 billion triples in July 2009 (Bizer, 2009)), the specific data required to annotate a given collection may not be available in a single pre-existing resource. However, this data may be publicly available in some other form on the Internet: possibly published in a database, accessed via some query form, embedded in some semi-structured html, as lists or tables, or simply present within some free-text. In this paper we consider the use of *focused crawling* (Chakrabarti et al., 1999) and automatic *information extraction* (Cowie and Lehnert, 1996) techniques to exploit Web pages for a semantic annotation task. The annotation process initially uses domain-specific information sources, but supplements these with information gathered from more general resources to provide the necessary coverage as required.

This paper expands upon previous work (Clough et al., 2008) and is structured as follows: Section 2 discusses related work; in Section 3 we discuss a study undertaken with Tate Online to investigate potential requirements by current users for semantically-enriched access to Tate Online; in Section 4 we focus on the use of semantic annotation techniques to meet the requirements of cultural heritage information in the light of user's requirements; Section 5 discusses experimental evaluation of our proposed information extraction techniques for providing semantic enrichment to cultural objects based on the findings of the user study; and finally Section 6 provides a summary of our work to date and ideas for future directions.

2. Related Work

2.1. Semantic Web Technologies in the Cultural Heritage Domain

The principle of using the Semantic Web to integrate cultural heritage (CH) collections has been most recently expounded by Jankowski, et al. (2009) who discuss the benefits available to CH institutes if they design their applications according to the Linked Data principles. Hardman, et al., 2009 identified CH *"...information tasks that can be supported using linked data by making non-obvious connections among related pieces of information explicit, such as exploratory tasks or topic search."* Such work explores technologies to satisfy user needs, especially those of CH experts whose search tasks often involve relatively complex information gathering and use, combining results from multiple sources (Amin, et al., 2008). Two projects, eChase (Sinclair, et al., 2006) and AMA (Eide, et al., 2008), have developed tools which facilitate the translation of data into a Semantic Web format. A similar approach, the mapping of current metadata into a Linked Data format, was pursued in the MultimediaN E-Culture project (Ossenbruggen, et al., 2007), which also developed a prototype tool to show the potential of semantic access, including the use of faceted browsing and timelines. However, none of this work addresses the critical issue of acquiring the semantic information if it is not already available to the organisations in a well-structured form.

2.2. Information Extraction

Previous work has examine the used of information extraction from html (and XML) data, using tree-based wrappers (Ling, et al., 1999; Ling, et al., 2000; Baumgartner, et al., 2001; Muslea, et al., 1998) and in a number of cases the work has utilised supervised learning approaches (Cohen and Jensen, 2001; Sakamoto, et al., 2001). However, these systems rely on the labelling of the tree nodes to provide a training set for a supervised-learning approach. For any reasonably sized training set, such labelling requires time and expertise. The problem of acquiring these labels is addressed by the semi-supervised (bootstrapping) approach adopted in this paper. A similar approach to the one adopted here is the KNOWITALL system (Etzioni, 2004), however their approach searches from domain-independent information and, therefore, does not have to consider the issue of discovering information sources and of "out-of-domain" information, their system also relies on linguistic patterns, unlike our language agnostic approach.

The approach adopted in this work is most related to Knoblock et al's work, recently developed in the DEIMOS system (Ambite, et al., 2008). The main differentiation between the two systems is that DEIMOS uses query forms which generate webpages as information sources rather than webpages directly. To discover the information sources they use social bookmarking sites, such as del.icio.us., to provide links from known sources to others. DEIMOS then learns the mapping from query inputs to webpage outputs, employing a similar DOM tree extraction pattern Information Extraction approach to the one described in this paper. The use of such automatically generated webpages means they are more likely to have regularity in the html produced, which should improve extraction precision, however the requirement for a form interface limits the information sources and thus is likely to adversely affect recall.

An interesting analogy can also be made to the Price Comparison systems, initially these used webpage scraping techniques but now rely more on feeds of information from the retailers or third parties. It is possible that if the large commercial websites (such as Artnet and World Wide Arts Resources) could see advantages in providing the artist information in a more structured form, either via an information feed directly to their database or using more strict markup, by applying microformats (Khare and Çelik, 2006) to specifically indicate a value's semantic information type within their webpages, then the process of information extraction would be greatly simplified. It is likely that such sites would require some form of monetary gain from providing this semantic annotation, however it is possible that providing this information to organisation such as the Tate may benefit the Tate users, in browsing and searching the collection, and in turn lead those users to the commercial linked data source.

3. A Case Study: Tate Online

In previous work we explored how Semantic Web technologies could be applied to enhance information access to Tate Online (Clough et al., 2008). The Tate is Britain's national gallery and houses both the national collection of British art from the 16th Century and international modern art. The Tate has one collection shared between four physical galleries: Tate Britain, Tate Modern, Tate Liverpool, and Tate St Ives. Each gallery has an online presence drawing from the same database or collection of digitised content (over 65,000 works in the collection). The Tate Modern Gallery is the world's most visited museum of contemporary art, with over 4 million visitors per year. Similarly, the Tate Online website attracted over 7 million unique visitors in 2005. It is one of the most popular UK visual arts and museum websites. Given the international importance of content provided by the Tate Gallery, semantic search would seem an ideal way in which to increase accessibility to the online collections, and thereby increase traffic to the website. Tate Online provided us a unique case study in which to examine the utility and feasibility of utilising technologies for enhancing search.

3.1. Users' Requirements for Semantic Search

As an initial means of gathering background information on users' typical tasks and needs when using a site like Tate Online, a questionnaire was published targeting visitors of the site. This survey was offered as a pop-up to individuals visiting the Tate Collection website. It was conducted to get an idea of what people use the Collection site for and to use this input to help guide the design of a system for browsing and exploring material related to artists and artworks. A total of 635 individuals world-wide answered the online questionnaire. Of these responses, 42% stated their primary reasons for visiting the Tate site were related to academic/research objectives, and 34% were using it out of personal interest. Roughly 2/3 of people visited the site looking for something specific, such as a particular artist (45%) or artwork (19%), or both (12%). Alternatively, 14% looked for types of artworks, and 10% were just browsing the collection. When asked which criteria would be the most important when searching for an artwork or artist (from the list in Table 1), overall 45% of respondents mentioned subjects of artworks, 31% voted for relationships between artists, and 13% selected dates (such as artists' birth dates or artwork creation dates). Table 1 shows similarities between user groups (general and expert user): although absolute percentages vary, the ranking of functionalities remains the same.

Table 1. Most important criterion for search, by user type

Topic	Total %	% of general users	% of expert users
Artwork subjects	44.7	34.6	51.5
Relationships	31.1	38.8	23.3
Dates	13.1	13.5	15.2
Gender of artists	3.5	13.5	4.0
Nationality of artists	3.6	7.7	3.0
Locations	4.0	1.9	3.0

Respondents were also asked which of a range of possible features would be most useful in enhancing access to material in the Tate Collection. The "most useful" feature as chosen by the greatest number of respondents (26.2%) was faceted browsing (the ability to search for information based on several criteria at once, e.g. "find female French artists from the 19th century"). Also of high interest was the ability to explore relationships between artists. Finally, the possibility of accessing other (related) links in English via the Collection pages was deemed to be useful. This also emerged as a theme in a previous 2004 internal Tate Online survey, in which people expressed a wish to be able to access links to other sites (artists' official pages, other high quality art/museum webpages) provided on the Tate pages, in order to further their information seeking and exploring process. Upon investigation, the answers were similar between expert and general users; the most notable difference being the percentage of people who would find it useful to explore relationships between creators and creations: this was highly ranked by the expert users, but of low importance to the general users (for whom links in English were more important).

3.2. Prototype Development

Based on previous findings and the availability of (semantic) data through information enrichment (Section 4), we developed an online prototype to demonstrate the potential of semantics in accessing and exploring the Tate Collection (see Clough et al., 2008). With the permission of the Tate we scraped the Tate Online website and created a number of "wrappers" to extract the structured data used to generate the pages. This provided information on the (approximately 3,000) artists on the site (i.e. their names, birth/death dates), and information and images on their (approximately 30,000) related artworks (i.e. titles, subject). This information was augmented by linking the Tate artists to the information provided by Getty Union List of Artist Names (ULAN). ULAN contains information on over 100,000 artists, including name variations, nationality, birth and death dates/places, role, gender, relationships: these form the facets for exploration.

Faceted browsing has been shown an effective means of supporting exploratory search (Hearst, 2006), as well as well as helping users find a specific type of information or to help narrow down a search (Capra et al., 2007). The prototype system provided access to the same collection via a faceted browsing interface. A number of faceted browsing development tools are available (e.g. SlashFacet, mSpace, Flamenco) and the prototype system was implemented using the Simile Exhibit toolkit from MIT. This offers a lightweight (it is implemented entirely in JavaScript and the interface is configurable via the webpage html) and comprehensive system (including a variety of types of facet (i.e. numerical, hierarchical) and views (i.e. timeline, maps)), which allows for flexible and fast prototyping.

4. Semantic Annotation

Semantic Annotation is used both with the term 'annotation' in its noun form, to describe the annotations assigned to an instance of a concept, and verb form, the process of assigning the annotations. For example, the instance of the Artist concept related to Pablo Picasso can have the semantic annotations Name "Pablo Picasso" and BirthDate "25th October 1881", and these may have been derived from the process of semantic annotation of the text strings found in some database, document, webpage, etc. In the current study, we examine the semantic annotation of the artists found in the Collections section of Tate Online. The process involves associating the instances of artists to those found in a structured domain-specific resource, in this case ULAN (see Section 4.1). The coverage of this resource is extended to include artists and their related concepts appearing in data from the Tate, but missing from ULAN. This is achieved using data mined from unstructured (and semi-structured) online information sources with a semantic annotation technique and forms the main focus on this paper. Section 4.1 discusses various sources of information pertaining to artists, both structured and unstructured; Section 4.2 highlights one of the main issues in using such information: dealing with imperfections; Section 4.3 discusses our approach for extending ULAN by mining data from various online sources using focused crawling and information extraction.

4.1. Information Sources

In order to acquire the related concepts (or properties) associated with an artist, one of the primary sources of publicly-accessible information in the cultural heritage domain is the Getty Union List of Artist Names (ULAN). This resource provides structured information on over 116544 artists, where each artist can have a number of associated properties, such as:

- Preferred Term (i.e. name)
- NonPreferred Terms (i.e. alternative names)
- Birth/Death Date

- **Birth/Death Place**
- **Nationality**
- **Role (e.g. painter, sculptor)**
- **Gender**
- **Associated (typed) Relationships links to other artists in ULAN (e.g. teacher_of, married_to)**

Although relationships between different artists were highlighted as important by users of Tate Online (Section 3.1), it was not included as a concept for semantic annotations as initial studies showed that such information was very sparsely represented. Therefore, the following work considers the annotation of seven semantic concepts: BirthDate, DeathDate, BirthPlace, DeathPlace, Nationality, Role and Gender. Whilst ULAN does provide a well-researched and extensive resource, it also has a number of limitations:

- **It is expensive to create and maintain (therefore it is not free, but requires an annual license).**
- **Whilst it is well-researched, it still contains errors (although each release of ULAN both adds to the number of artist and their properties and corrects the discovered errors).**
- **For most of the properties ULAN provides a single value, however it is possible that the actual property value for a given artist is not precisely known or is under debate. Therefore ULAN contains an editorial bias which may not reflect the nuances of the domain information (i.e. the diversity of opinions or true nature of information imperfection) or the particular bias of the users who require that information.**
- **Whilst it is extensive the resource is not complete, both in terms of its coverage of artists and the properties associated with those artists. The Getty Institute, which collects the artist information, may exhibit a cultural bias in the sources of research they use to collect the artist information which limits the likelihood of certain artists appearing on the list.**

Therefore for a given user (i.e. cultural heritage institution), ULAN may be more or less applicable to their requirements. For example, when considering the coverage between the Tate Collection and ULAN, 73% of the artists in the Tate Collection are covered by ULAN (see Table 4 below), which leaves over a quarter of the artists for which the extended properties of ULAN are not available. To mitigate the limitations of ULAN, we evaluate the World-Wide Web (WWW) as a potential source of information.

The WWW contains vast amounts of information: current estimates of the number of visible indexed webpages totals around 25 billion (WorldWideWebSize), and it is increasingly seen as a repository of collective intelligence. However, finding the required information can be problematic: the information presented on the WWW covers almost every known subject domain (and sub-domain), may be of unreliable quality and distributed across a number of webpages. Therefore, techniques employing the WWW as an information source must discover and combine the required information and then provide some assessment of its quality (i.e. likelihood of being correct). In utilising the WWW, the approach adopted in this work makes certain assumptions about the nature of the information sources in the domain of interest:

- **The first assumption is that the information on a page is likely to relate to a given instance of a concept if the title of the webpage contains distinguishing properties of that instance. With the caveat that unless the property value is a unique identifier (such as an ISBN number) the page may relate to a different entity (i.e. artist) which possesses the same value (i.e. name). A further disambiguation process may then be required to ensure that page relates to the desired entity. This assumption is necessary to assign the information extracted from the webpage to the page's concept.**

- In addition, it is assumed that information from a given web site is likely to relate to a domain if the webpages in that site relate to instances of domain concepts, and tautologically that webpages are more likely to relate to domain concepts if they are within a web site which is deemed to relate to the domain. With the caveat that a given web site may relate to a wider domain than the domain of interest.
- It is assumed that pages from the same website will contain syntactic regularity that mirrors semantics. This assumption will allow effective structural extraction patterns to be formed.
- Finally, it is assumed that related information from different websites is independent. This redundancy enables the extracted information to be validated, due to the discovery of multiple occurrences of concept property values.

The first assumption will limit the amount of information which can be extracted from the WWW as webpages which present information about multiple artists on a single page are excluded. However, requiring that a webpage be about a single artist increases the likelihood that any information extracted from the page can be assigned to that artist. In order to compare the WWW with ULAN as an information source, we search the WWW for webpages whose title contains the ULAN artist names using Google's Search API (GoogleAPI). At most, the top 10 search results for each artist name were crawled, this resulted in the retrieval of 812,327 webpages, from 7124 websites. The retrieved pages were then searched for the selected property values associated with that ULAN artist. The tables below show the results of this process (direct references to the ULAN data, such as the Getty website, were excluded from the search). Table 2a shows the extent to which the individual artist properties were retrieved and Table 2b shows the number of properties for each artist which were retrieved, both considering properties on the single most informative webpage and also the properties on all the webpages which relate to that artist.

From Table 2a, 80% of ULAN artists potentially have some equivalent webpage, although only 62% of artists have webpages which contain at least one ULAN property. Thus 20% at least of the ULAN artists do not appear to have a related webpage, however it is worth noting that on websites such as Artnet, whilst 26575 ULAN artist webpages were identified, the website claims that pages on over 180000 artists are available. Similarly, on the World Wide Arts Resources (WWAR) website, 6657 ULAN artist webpages were retrieved from a total of 22000 artists' pages. Therefore, although only a proportion of the ULAN artists have relevant webpages, there are apparently a significant number of webpages for artists which are not covered in ULAN. Similarly, Table 2b shows the proportion of ULAN properties contained within each artist's page. However, it is possible that the property values missing from ULAN may be available from these pages, and thus Information Extraction methods (discussed below) may be able to extract those values.

In Table 2a the difference between the properties contained in the single most informative webpage and the properties combined from all the webpages indicates that it is necessary to retrieve and combine multiple sources of information (webpages) for each artist to extract the maximum number of available properties.

Table 2a. ULAN artist covered by (the single most informative, and all) WWW artist pages

	Minimum number of ULAN artist properties contained in the webpage							
	>=0	>=1	>=2	>=3	>=4	>=5	>=6	>=7
Retrieved (single webpage)	93074	72624	48173	30015	12470	4864	1796	75
Retrieved (all webpage)	93074	72624	53321	35373	17230	7977	3460	483

Total	116544	116541	116537	111698	86340	44471	17404	10867
Recall (single webpage)	0.8	0.62	0.41	0.26	0.14	0.1	0.1	0.01
Recall (all webpage)	0.8	0.62	0.45	0.31	0.19	0.17	0.19	0.04

Table 2b. ULAN artist properties covered by WWW artist pages

	Artist Properties						
	BirthDate	DeathDate	BirthPlace	DeathPlace	Nationality	Role	Gender
Retrieved	50583	35648	10956	7237	55789	17894	12361
Total	109140	53928	23611	14015	116541	70111	116512
Recall	0.46	0.66	0.46	0.52	0.48	0.26	0.11

One of the most common sources of artist webpages is Wikipedia (wikipedia.org). The English Wikipedia pages (en.wikipedia.org) cover 16356 ULAN artists, while from all the Wikipedia pages (55 different Wikipedia languages sites) 21140 ULAN artists are covered. Wikipedia has been shown to be at least comparable to a traditional encyclopedia in terms of accuracy (Giles, 2005), and this, together with the fact it is publicly and freely accessible, has led to a considerable amount of work on extracting structured data from Wikipedia articles. Mike Bergman (Bergman, 2009) provides a comprehensive list of references for this work. The most complete, and widely used, extracted data is provided by DBPedia which applies hand-coded scripts to extract semantic data from the semi-structured Wikipedia articles. Table 3a shows the ULAN artists covered by DBPedia and Table 3b the properties available for those artists.

Table 3a. ULAN artist covered by DBPedia artist resources

	Minimum number of ULAN artist properties contained in the DBPedia resource							
	>=0	>=1	>=2	>=3	>=4	>=5	>=6	>=7
Retrieved	14972	11635	8778	6757	4455	2464	1381	357
Total	116544	116541	116537	111698	86340	44471	17404	10867
Recall	0.12	0.09	0.07	0.06	0.05	0.05	0.07	0.03

Table 3b. ULAN artist properties covered by DBPedia artist resource

	Artist Properties						
	BirthDate	DeathDate	BirthPlace	DeathPlace	Nationality	Role	Gender
Retrieved	8077	6226	3002	2069	9628	5442	1383
Total	109140	53928	23611	14015	116541	70111	116512
Recall	0.07	0.12	0.13	0.15	0.08	0.08	0.01

Table 4 shows the combined coverage of artists referenced in the Tate Collection section of Tate Online by ULAN, Wikipedia and general webpages. It should be noted that the figures relate to all webpages and not those that contain artist properties, thus pages may refer to entities other than the artist. Wikipedia is divided into English (EN), non-English (non-EN) and all pages (All) and the WWW is divided into all pages (All) pages from web sites which contain at least 2 webpages with the names of Tate artists (2+) and at least 10 webpages with Tate artist names (10+).

The diagonal in the table indicates the extent to which each of the resources covers the artists in the Tate Collection. The other values indicate the extent to which the combined resources cover artists in the Tate Collection. As can be seen from the results, ULAN covers 73% of the Tate Collection, Wikipedia covers 55.4% and the WWW covers 99.5%. The table indicates that there is potential benefit in using the combination of resources to discover and extract information. In fact Wikipedia and ULAN in combination cover 82% of the Tate Collection which shows that Wikipedia (in the form of its associated DBPedia semantic data) could be used as a general structured resource to augment the domain resource.

Table 4. Combined Coverage of the Tate artists in ULAN, Wikipedia and the WWW

		ULAN	Wikipedia			WWW		
			EN	nonEN	All	All	2+	10+
ULAN		0.730	0.800	0.783	0.820	0.996	0.982	0.968
Wikipedia	EN		0.496	0.554	0.554	0.995	0.980	0.961
	Non EN			0.339			0.972	0.947
	All						0.982	0.967
WWW	All					0.995		
	2+						0.962	
	10+							0.924

4.2. Recognising Imperfections in the Information

In many information systems it is necessary to deal with the issue of imperfect information, i.e. any situation where complete and accurate information is not available. In the Cultural Heritage domain, where much of the information is from multiple (and potentially conflicting), distributed, biased and historic sources, imperfect information is pervasive. The nature of the such imperfections can take a number of forms, including:

- *Missing*: the most basic form of imperfect information is when it is unavailable. However it may be that the information is only partially missing, for example where we know some of the properties associated with an artist, or when a set of correct values is possible (as is the case with artist roles) and only a subset of those values is known.
- *Uncertain*: where the accuracy of a value is not known or is in debate, in some cases it may be possible to place some measure of the uncertainty in a piece of information.
- *Imprecise*: where the information is correct but to varying degrees of precision (e.g. birth year/range/date, location country/region/city).
- *Ambiguous*: where multiple correct values are possible (i.e. artists known by multiple names). However the correctness of a particular value may depend on the context of its use, for example where we are dealing with multilingual data, a correct piece of information in the wrong language may not satisfy the information requirements.

To an extent these imperfect forms are interrelated. For example, we may be certain about an artists' year of birth, but uncertain about the actual date. The reason for the existence of imperfect information is imperfections in the collection, storage or retrieval of information. Where there is a lack of proximity (e.g. in time or distance) between an event or object and the recording of the information relating to that event or object, errors may occur. The recording, or copying of information may involve bias. In some cases the "true" information may not be available, academic debate may provide some agreed upon value (i.e. truth by consensus) or an approximate, range or set of values may be as close to the truth as is possible.

In this work two "Gold Standard" sources are used: ULAN and the Tate. Looking at the common properties of these two source (i.e. birth/death) dates there is a degree of disagreement (Reference Interannotator disagreement issues), which needs to be considered when using the data. Values from the Tate Collection, whilst accurate, should not be seen as precluding other values from being taken as correct. ULAN, as a widely used and well respected domain resource, can also be seen as providing Gold Standard data. However, it is interesting to note that for comparable date values between the Tate and ULAN, there is a degree of disagreement. Figures 1 and 2 show the frequency distributions of differences between birth/death year for the same artist from the Tate and ULAN.

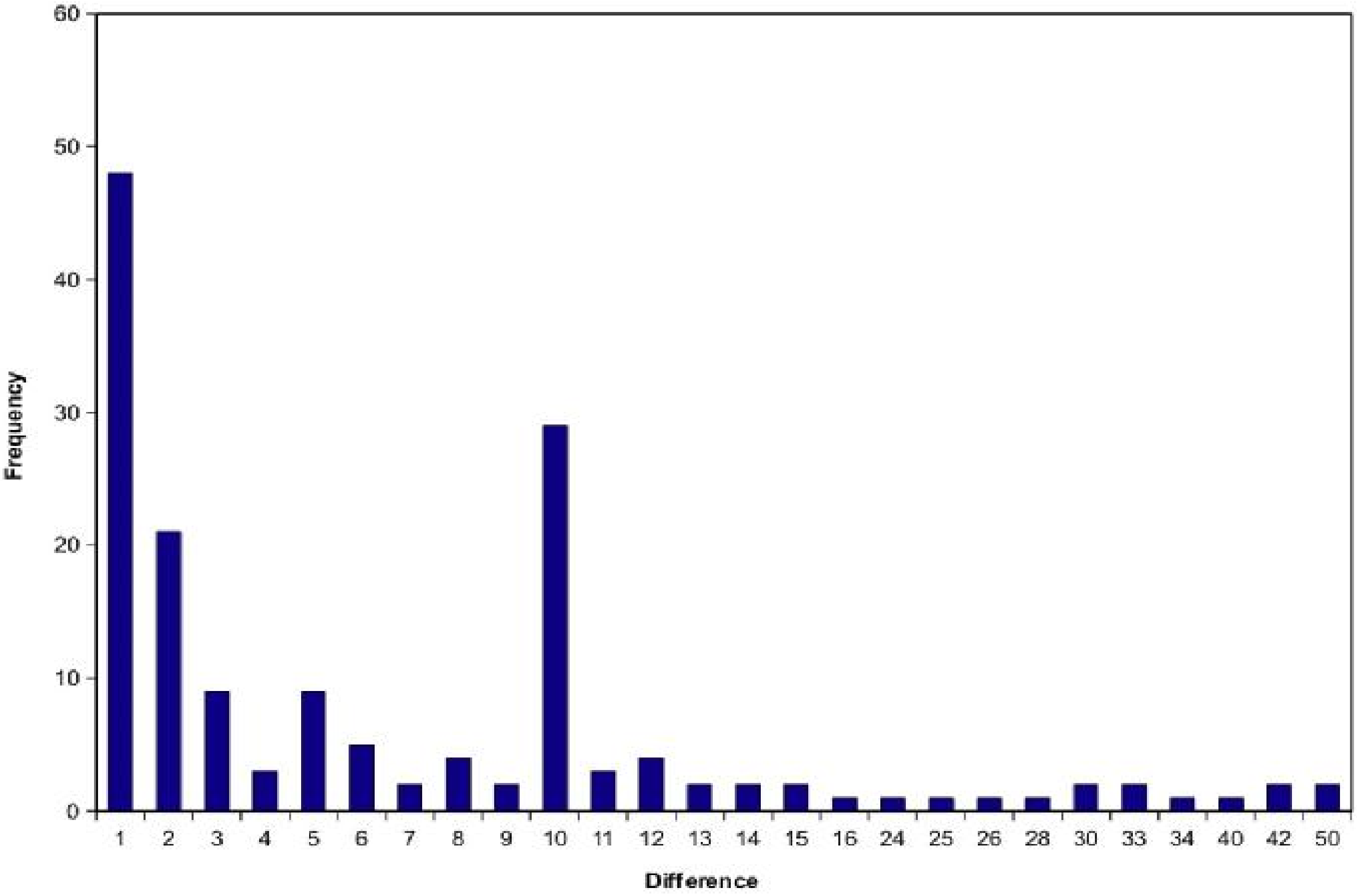

Figure 1. Tate/ULAN Artist Birth Date Differences

There are 2088 artists who have a birth date in both the Tate Collection and ULAN. Figure 1 shows the 160 (7.7%) of artists for which there is a disagreement in the date of birth. Of these 43.1% disagree by only 1 or 2 years, however the disagreement ranges up to 50 years. There is a noticeable peak at 10 years which is due to the case where there is limited precision for the birth dates and the nearest decade is selected (with Tate and ULAN researchers disagreeing on the chosen decade).

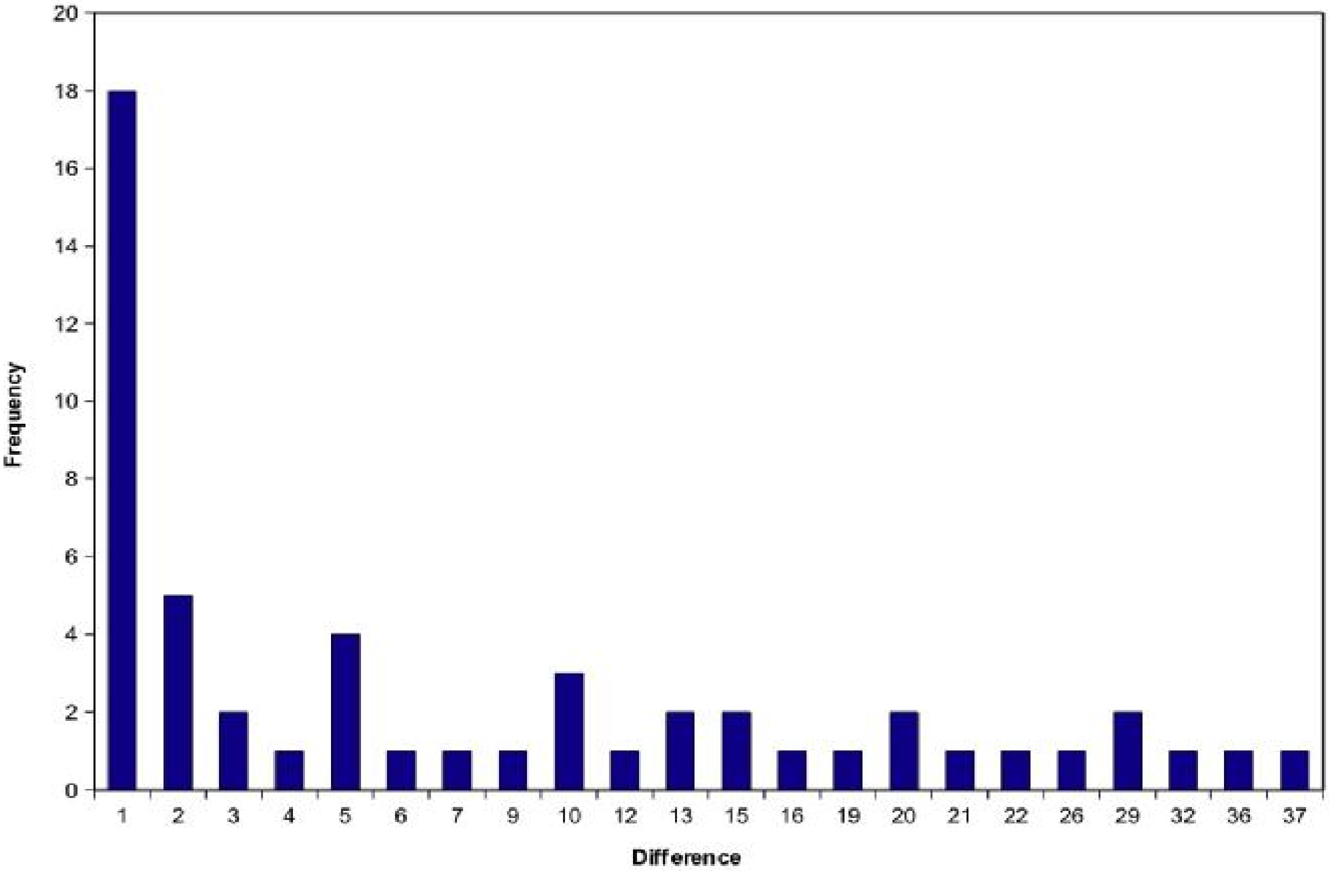

Figure 2. Tate / ULAN Artist Death Date Differences

Artist death dates are less prevalent than birth dates, both as they are seen as of secondary importance and because some artists are still living. Thus there are only 1345 artists who have a death date in both the Tate Collection and ULAN. Figure 1 shows the 53 (3.9%) for which there is a disagreement, with 43.4% of these disagreeing by only 1 or 2 years.

In addition, whilst processing the ULAN data a number of incorrect values were discovered. Some of these are strange systematic anomalies (e.g. there is always a value for the date of birth/death, even where this is unknown or inappropriate), whilst other information was obviously incorrect. These anomalies are simply removed by ensuring that values adhere to the domain ontology constraints (e.g. maximum age is 100, birth/death dates are less than the creation date of the information source, people cannot be married to themselves, etc.).

4.3. Extending Pre-Existing Sources with Information Extraction

Information extraction can be seen as a process for extracting structured information from unstructured data, or alternatively making explicit the information which is implicit in the representation of a given resource. Therefore, the appropriate information extraction technique to employ depends upon the resource's representation. In order to extract information from webpages it is necessary to exploit *redundancy* and *regularity*. Redundancy on the WWW means that information tends to be repeated on a number of different sources (webpages), therefore multiple occurrences of the same piece of information increases the likelihood of that information's correctness, due to the combined evidence. To make use of this fact there must be an assumption that the different information sources are, to some extent, independent. Ideally the extraction process should consider the degree to which this assumption is valid for a given piece of information.

Regularity means that information tends to be represented in a similar format (syntax). This regularity is likely to be more pronounced when the information comes from a similar source (i.e. webpages from the same website), as a uniform format is influenced by style, editorial preferences and guidelines. Information extraction from free-text has to rely on regular patterns of grammatical syntax. In order to exploit this syntax to discover regular patterns it is necessary to normalise the text using language specific techniques, such as tokenisation, stemming, lemmatisation, chunking, parsing, etc. However, a desirable feature of our approach is that it is applicable to multiple language sources, and thus a language "agnostic" extraction process is employed, which utilises the formatting structure of the webpage. This language independent approach means the extraction process is not reliant on linguistic patterns in the text and thus can extract information which is defined by its position in the page's structure, such as in tables, lists, headings, etc.

The textual information sources (i.e. webpages) used by this process wrap text in Hyper-Text Markup Language (html). The primary purpose of html is to format the display of information for human readability. However this often means that for well structured pages the values for semantic concepts (i.e. artist properties) are contained in separate markup, so that they can be emphasised or linked to other pages relating to that specific concept. For some websites information is derived from databases and displayed using standard style sheets, such information can be extremely regular. The Information Extraction process exploits the html representation by transforming the webpage into its Document Object Model (DOM) which provides a tree structure where the text on the webpage is contained within the various nodes in the tree. The extraction process attempts to determine the nodes which contain specific semantic data (i.e. artist properties). The extraction patterns, which are represented as the paths from the root of the DOM to the semantic nodes, are then used to extract information from other pages on the same website. The process is as follows:

1. The process is initiated with a set of "seed" examples of artists and their properties (from ULAN).
2. The artist name is used as the search term with which to located potential relevant webpages by retrieving pages which contain the name in the title text.
3. All text is normalised into an ASCII representation (to remove any diacritics and ligatures which are not uniformly used).
4. Those DOM nodes that contain text which matches property values (either exactly or approximately) and do not contain conflicting values are annotated as semantic nodes.
5. Generate extraction patterns, i.e. path from the root node to the semantic nodes.
6. Apply the none conflicting extraction patterns to extract information to other pages on the same website.
7. Evaluate the extraction patterns, webpages and extracted information (see below).

For the known (seed) examples the evaluation of the extraction pattern depends upon the degree to which the property value contained in the semantic node is exact and unambiguous. For example, from the previous Section 4.2 on Information Imperfection, if the node contains dates are within 10 years of the ULAN (birth/death) date, then the dates are deemed to match. Similarly, if the node contains a place (e.g. England) which contains the ULAN (birth/death) place (e.g. London), then the places are deemed to match. However, the text in each semantic node is also searched for semantic values other than those in the seed example (i.e. other dates, places, nationality, roles and genders). For dates (values from 1000 to 2007) and genders this is straightforward, however for the other semantic concepts domain gazetteers are required. For references to place, the Getty Thesaurus of Geographic Names (TGN) is used (ULAN place properties reference the TGN place values). For nationalities and roles all the values found in ULAN are used to create the gazetteers.

Given that a node in the DOM contains a known property value for that artists, that node is labelled as a "semantic node" and used to form an extraction pattern unless it contains conflicting values. The nature of conflicting values depends upon the semantic concept:

- *Dates*: if the semantic nodes contain dates below for a BirthDate node, or dates above for a DeathDate node.
- *Places*: if the semantic nodes contain other places, aside from places which either contain or are contained within the birth/death place. For example, if a semantic node contains the known BirthPlace "South Yorkshire", it is considered valid even if it also contains the places "England" and "Sheffield".
- *Nationality:* if the semantic nodes contain any other nationality.
- *Roles*: if the semantic nodes contain any other role. This is a very strict rule as artists can have multiple roles, however the desire is to create high precision extraction patterns.
- *Genders*: if the semantic nodes contain any other gender.

In addition in order to be retained the extraction patterns are required to extract at least two different semantic values, i.e. a BirthDate extraction pattern must extract two different dates. This increases the likelihood that the semantic node does indeed relate to the semantic concept, rather than some spuriously matched value. After the inexact and inaccurate extraction patterns have been removed, the resultant set of patterns are applied to the other artist webpages. Note that the extraction patterns are website specific. This results in a set of semantic values extracted for each artist. The degree to which a value is determined to be correct is a function of: the quality of the extraction pattern (i.e. the degree to which it extracts precise and unambiguous values) and the uniformity of the extracted value set.

5. Evaluation

5.1. Evaluating Semantic Annotation

In order to determine the extent to which the semantic annotation process discussed in the section above is able to extract accurate information, it is necessary to determine the degree to which the size of the seed example set influences the performance of the Information Extraction technique. A 5-fold cross-validation experiment was performed using the ULAN data. The following set of figures shows the Recall (Figure 3), Precision (Figure 4) and F-Measure (the harmonic mean of Precision and Recall values, see Figure 5) for the seven artist properties, for a varying degree of seed example size. A value is deemed to be correctly extracted if the known value is in the set of extracted values for an artist's concept.

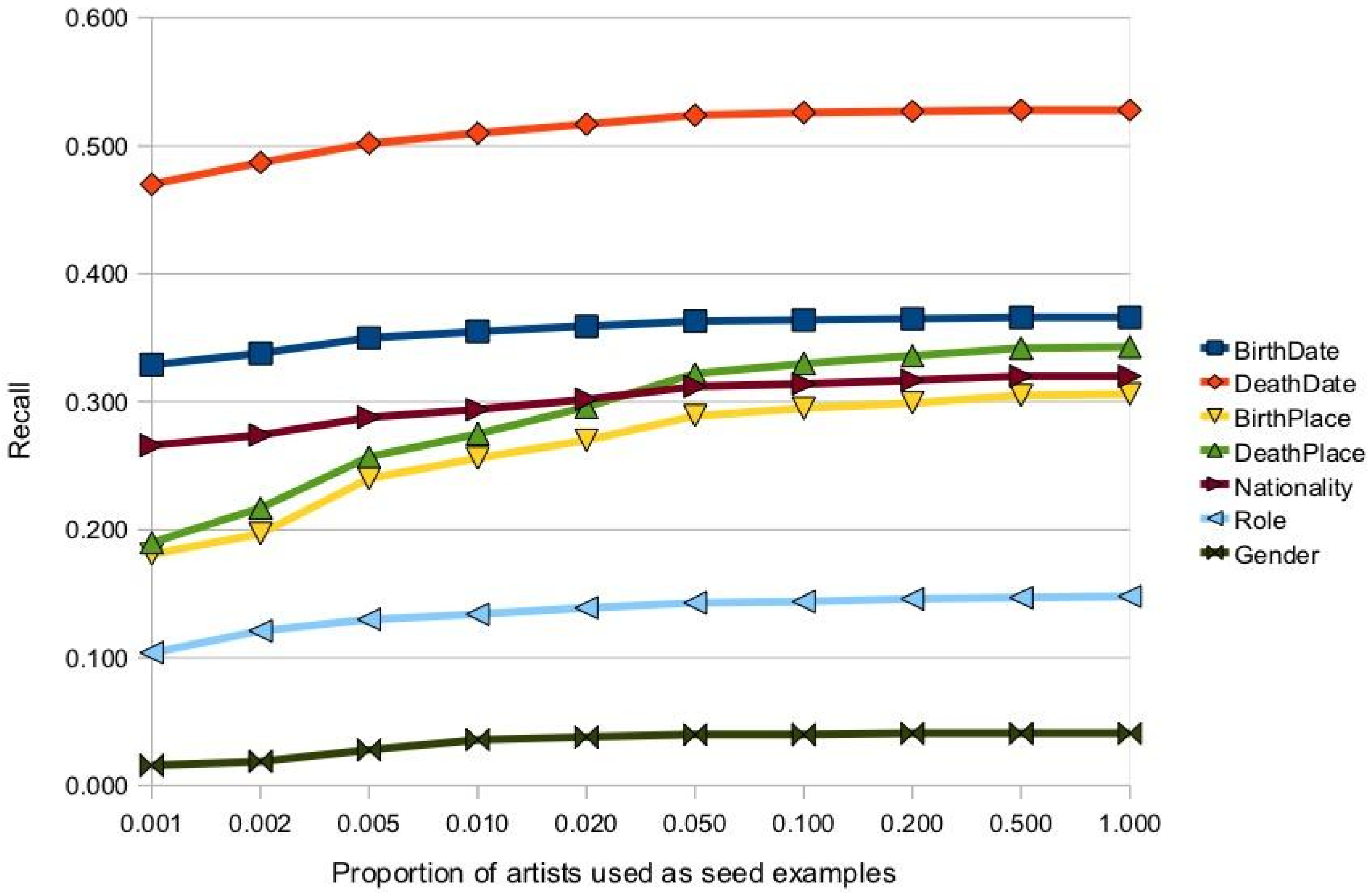

Figure 3. ULAN Artist Recall, varying seed example set proportion

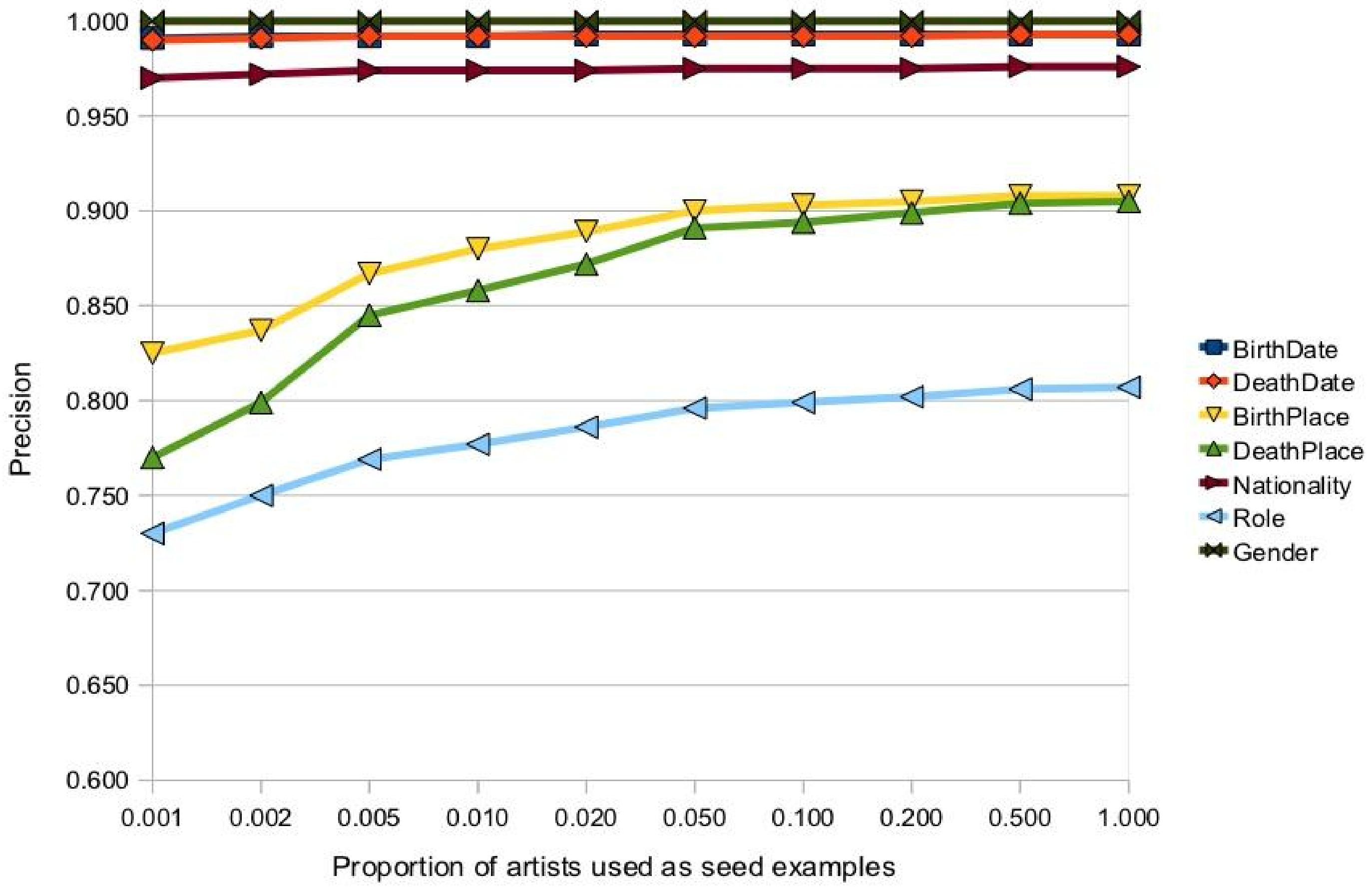

Figure 4. ULAN Artist Precision, varying seed example set proportion

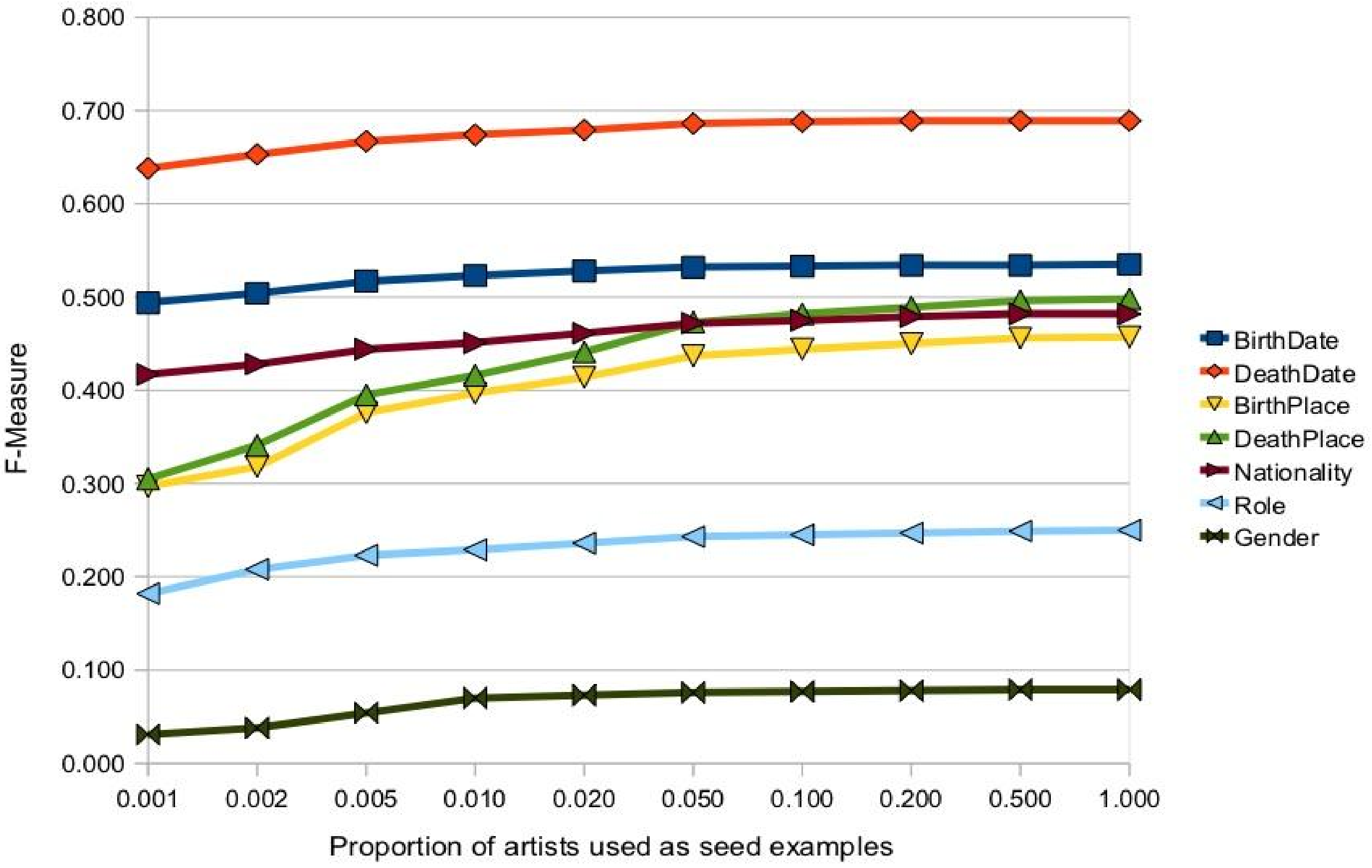

Figure 5. ULAN Artist F-Measure, varying seed example set proportion

Figure 3 shows that there is a wide variation in recall for the artist properties, with artists' roles and, especially, genders being poorly represented on webpages, in comparison with ULAN. It is possible that the webpage creators do not consider such properties important, as gender is largely obvious and assigning roles more specific than artist unnecessary. Another interesting feature of the graph is that the death related information has a higher recall than the birth related information. This is possibly due to the fact that, as it is not deemed necessary to include death related information, it is only included when definitely known and is therefore more consistently represented for a given artist.

Figure 4 shows that Precision is high for properties, except the location properties (birth/death place) and role. These properties are also most affected by the seed example set size. One possible explanation for this is that these properties are seen as of secondary importance behind artists' dates and therefore are less likely to have a standard representation on the webpage. It is more likely to be retrieved from passages of text where there is an increase likelihood of an ambiguous extraction of values.

The graphs show that the performance does not drastically decrease, even for very low seed examples set size (0.001 is equivalent to around 100 artists). This is due to the fact that around half the ULAN artists (55014) are represented on 248 of the most informative websites. Therefore, any reasonable number of seed examples will provide access to the information on these sites.

5.2. Semantic Annotation of the Tate Collection

In the final experiment the semantic annotation approach was applied to the 3000 artists present in the Tate Collection data, shown in Table 5, showing both absolute and percentage coverage of Tate artists' properties. Note that only 2190 Tate artists are also present in ULAN. This indicates that the Information Extraction techniques complement the ULAN information, most notably for date and location information, with the caveat that location properties have a relatively low precision when the number of seed examples is low, although their precision may be as high as 90% (see Figure 4). Thus, for the 3000 artists in the Tate Collection, ULAN covers 45.1% of all the possible values. However, by utilising Information Extraction of values from the WWW 63.3% of all the values are covered (an improvement of 18.2%).

Table 5. Artist Properties Retrieved from ULAN and the WWW for the Tate Collection

	Artist Properties						
	BirthDate	DeathDate	BirthPlace	DeathPlace	Nationality	Role	Gender
ULAN	2051 (63.4%)	1013 (33.8%)	444 (14.8%)	263 (8.8%)	2190 (73%)	1317 (43.9%)	2189 (73%)
WWW	379 (12.6%)	1017 (33.9%)	868 (28.9%)	986 (32.9%)	262 (8.7%)	281 (9.4%)	35 (1.2%)
Total	2430 (81%)	2031 (67.7%)	1312 (43.7%)	1249 (41.6%)	2452 (81.7%)	1599 (53.3%)	2224 (74.1%)

6. Conclusions and Future Work

It is clear that technologies from the Semantic Web have the potential to improve information access to cultural heritage collections. As the availability of publicly-accessible data in an interoperable form (e.g. as linked data) increases, the potential for linking and sharing cultural heritage material with other resources also increases. Not only may this help raise the visibility of content held by a provider, but it may also benefit the end user in providing a wider context for knowledge discovery and exploration. Technologies for semantic annotation can be used to make the semantics of digital content explicit automatically from cultural heritage material and thereby support the linking to pre-existing information sources.

In this paper we have explored the use of semantic annotation in a specific case study automatic annotation of information about artists from Tate Online using an existing structured and generally-accessible resource, the Union List of Artist Names (ULAN). By mapping instances of artists from content in the Tate Collection website to ULAN enables the provision of more sophisticated forms of information access, which are potentially better suited for users of humanities content, e.g. faceted search and browsing functionalities. However, the main problem with this approach (and any approach that uses pre-existing sources) is a lack in coverage of artists and their associated semantic properties (e.g. birth/death date, birth/death location) with respect to a given collection. Although this problem is a potential barrier to successful utilisation of semantic annotation techniques, it has received little attention in past work. It is this problem which forms the focus of our current investigations.

In this paper we investigate the success of extending the coverage of ULAN using focused crawling and automatic information extraction techniques to exploit semi-structured online sources of information (e.g. webpages and Wikipedia articles). Without extending ULAN we are able to provide values for 45.1% of the seven semantic concepts for artists represented in the Tate's data, by extending ULAN with information gathered online we are able to increase this to cover 63.3% of the values. In general, we demonstrate that semi-structured online sources can be utilized to increase the coverage of structured cultural heritage resources and could be applied to resources in other collections and domains.

However, we recognise that our current study is limited in at least three ways: Firstly, we have investigated only artist names and associated properties and not the relationships between artists and other named entities (or concepts), even though users of Tate Online indicated this as an important function to support. The main reason for this is that to establish relationships between named entities, one must first identify the named entities themselves and extend pre-existing sources to cover specific collections where necessary. Secondly, we recognise that the results from our study are based upon a single collection and scenario: artist names from Tate Online. Thirdly, although the information extraction technique employed is language agnostic, the gazetteers for the concept values are in English. Therefore, only information from non-English resources, such as dates (years) and names (if proper nouns are not translated) would be extracted.

In future work we plan to consider further case studies and also explore providing some of the other functionality indicated as useful for information access, such as relationships and automatic classification of artworks by subject. Analysis of the extent to which non-English webpages provide relevant information will also be explored.

4. Acknowledgments

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST-033104. We also thank the Tate for allowing us access to their data and other resources. The authors are solely responsible for the content of this paper.

5. References

- Ambite, J.L., Knoblock, C.A., Lerman, K., Plangprasopchok, A., Russ, T., Gazen, C., Minton, S. and Carman, M. (2008) "Exploiting Data Semantics to Discover, Extract, and Model Web Sources". In *Proceedings of the 2008 IEEE international Conference on Data Mining Workshops* (December 15 - 19, 2008). ICDMW. IEEE Computer Society, Washington, DC, pp. 771-779.
- Amin, A., van Ossenbruggen, J., Hardman, L., and van Nispen, A. (2008) "Understanding cultural heritage experts' information seeking needs". In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (Pittsburgh PA, PA, USA, June 16 - 20, 2008). JCDL '08. ACM, New York, NY, pp.39-47.
- Artnet. http://www.artnet.com (Accessed 1 December 2009).
- Baumgartner, R., Flesca, S. and Gottlob, G. (2001) "Supervised Wrapper Generation with Lixto". In *Proceedings of the 27th International Conference on Very Large Data Bases* (September 11 - 14, 2001). P. M. Apers, P. Atzeni, S. Ceri, S. Paraboschi, K. Ramamohanarao, and R. T. Snodgrass, Eds. Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, pp. 715-716.
- BBC Ontologies. http://www.bbc.co.uk/ontologies/ (Accessed 1 December 2009).
- Benjamins, V.R., Contreras, J., Blazquez, M., Dodero, J.M., Garcia, A., Navas, E., Hernandez, F., and Wert, C. (2004) "Cultural heritage and the semantic web". In *Proceedings of 1st European Semantic Web Symposium*, pp. 433-444.
- Bergman, M. (2009) 99 Wikipedia Sources Aiding the Semantic Web (Accessed 1 December 2009).

- Bizer, C. (2009) "The Emerging Web of Linked Data". *IEEE Intelligent Systems*, Vol. 24, No. 5, 87-92, Sep./Oct. 2009.
- Capra, R., Marchionini, G., Oh, J., Stutzman, F., and Zhang, Y. (2007) "Effects of structure and interaction style on distinct search tasks". In *Proceedings of 7th ACM/IEEE Joint Conference on Digital Libraries*, pp. 442-451.
- Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P. (2005) "Named graphs, provenance and trust". In *Proceedings of the 14th international Conference on World Wide Web* (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM, New York, NY, pp. 613-622.
- Chakrabarti, S., Berg, M., and Dom, B. (1999) "Focused crawling: A New Approach to Topic-Specific Web Resource Discovery". *Computer Networks: The International Journal of Computer and Telecommunications Networking* , Vol. 31, No. 11-16, 1623-1640.
- Cohen, W., and Jensen, L. (2001) "A structured wrapper induction system for extracting information from semi-structured documents." In *Proceedinngs of the IJCAI Workshop on Adaptive Text Extraction and Mining*.
- Cowie, J., and Lehnert, W. (1996) "Information Extraction". *Communications of the ACM*, Vol. 39, No. 1, 80-91.
- Clough, P., Marlow, J. and Ireson, N. (2008) "Enabling Semantic Access to Cultural Heritage: A Case Study of Tate Online". Larson, M., K. Fernie, J. Oomen and J. Cigarran (eds.) In *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage* , Aarhus, Denmark, September 18, 2008. ISBN 978-90-813489-1-1.http://ilps.science.uva.nl/IACH2008/proceedings/proceedings.html (Accessed 1 December 2009).
- DBPedia. http://dbpedia.org (Accessed 1 December 2009).
- Eide, Ø., Felicetti, A., Ore, C.E., D'Andre, A. and Holmen, J. (2008) "Encoding Cultural Heritage Information for the Semantic Web. Procedures for Data Integration through CIDOC-CRM Mapping". In *Proceedings of EPOCH Conference on Open Digital Cultural Heritage Systems*, pp. 1-7.
- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004) "Web-scale information extraction in KnowItAll: (preliminary results)". In *Proceedings of the 13th international Conference on World Wide Web* (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM, New York, NY, pp. 100-110.
- Fluit, C., Sabou, M. and Harmelen, F. van (2005) "Ontology-based information visualization: Towards Semantic Web applications". In V. Geroimenko (Ed.), Visualizing the Semantic Web (2nd ed.): Springer Verlag.
- Giles, J. (2005) "Internet encyclopedias go head to head". *Nature*, 438, 900-901.
- GoogleAPI. http://code.google.com/apis/ajaxsearch/web.html (Accessed 1 December 2009).
- Hardman, L., Ossenbruggen, J. van, Troncy, R., Amin, A., Hildebrand, H. (2009) "Interactive Information Access on the Web of Data". In *Proceedings of the WebSci'09: Society On-Line*, 18-20 March 2009, Athens, Greece. Available online:http://journal.webscience.org/212/ (Accessed 1 December 2009).
- Hearst, M. (2006) "Clustering versus faceted categories for information exploration". *Communications of the ACM*, Vol. 49, No. 4, 59-61.
- Hildebrand, M., Ossenbruggen, J. van, and Hardman, L. (2007) "An analysis of search-based user interaction on the semantic web". *Centrum voor Wiskunde en Informatica*.
- Hyvönen, E. (2007) "Semantic portals for cultural heritage". Available online:http://www.seco.tkk.fi/publications/2007/hyvonen-portals-2007.pdf (Accessed 1 December 2009).
- Jankowski, J., Cobos, Y., Hausenblas, M. and Decker, S. (2009) "Accessing Cultural Heritage using the Web of Data". In *Proceedings of 10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*, St.Julians, Malta, 2009.

- Khare, R. and Çelik, T. (2006) "Microformats: A Pragmatic Path to the Semantic Web". In *Poster Proceedings of the 15th International World Wide Web Conference*, Edinburgh, UK, May 2006. ACM Press, pp. 865-866.
- Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., and Lee, R. (2009) "Media Meets Semantic Web: How the BBC Uses DBpedia and Linked Data to Make Connections". In *Proceedings of the 6th European Semantic Web Conference on the Semantic Web: Research and Applications* (Heraklion, Crete, Greece, May 31 - June 04, 2009). L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds. Lecture Notes In Computer Science, vol. 5554. Springer-Verlag, Berlin, Heidelberg, pp. 723-737.
- Ling, L., Pu, C. and Han, W. (2000) "XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources". In *Proceedings of the 16th international Conference on Data Engineering* (February 28 - March 03, 2000). ICDE. IEEE Computer Society, Washington, DC, pp. 611-621.
- Ling, L., Han, W., Buttler, D., Pu, C. and Tang, W. (1999) "An XML-based Wrapper Generator for Web Information Extraction". In *Proceedings ACM SIGMOD International Conference on Management of Dat*a, June 1-3, 1999, Philadelphia, Pennsylvania, USA, pp. 540-543.
- Maedche, A. and Staab, S. (2002) "Applying semantic web technologies for tourism information systems". In*Proceedings of the Ninth International Conference for Information and Communication Technologies in Tourism (Enter-2002)*, pp. 311-319.
- Muslea, I., Minton, S. and Knoblock, C. (1998) "STALKER: Learning Extraction Rules for Semi-structured Web-based Information Sources." In Proceedings of AAAI-98 Workshop on AI and Information Integration, 74-81.
- Ossenbruggen, J. van, Amin, A., Hardman, L., Hildebrand, M., Assem, M. van, Omelayenko, B., Schreiber, G., Tordai, A., Boer, V. de, Wielinga, B., Wielemaker, J., Niet, M. de, Taekema, J., Orsouw, M.-F. van, and Teesing, A. (2007) "Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques". In *Proceedings of Museums and the Web 2007*, Toronto: Archives & Museum Informatics, published March 1, 2007. Available online:http://www.archimuse.com/mw2007/papers/ossenbruggen/ossenbruggen.html (Accessed 1 December 2009).
- Raimond, Y., Sutton, C. and Sandler, M. (2008) "Automatic interlinking of music datasets on the semantic web". In*Proceedings of the 1st Linked Data on the Web Workshop (LDOW2008)*.
- Sakamoto, H., Arimura, H., Arikawa, S. (2001) "Extracting Partial Structures from html Documents". In *Proceedings of the 14th International Florida Artificial Intelligence Research Symposium (FLAIRS2001): Knowledge Discovery and Data Mining*, AAAI Press, pp. 264-268.
- Sinclair, P., Lewis, P., Martinez, K., Addis, M., and Prideaux, D. (2006). "Semantic web integration of cultural heritage sources". In *Proceedings of the 15th international Conference on World Wide Web* (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM, New York, NY, pp. 1047-1048.
- TGN. Getty Thesaurus of Geographic Names. http://www.getty.edu/research/conducting_research/vocabularies/tgn/(Accessed 1 December 2009).
- ULAN. Unified List of Artist Names. http://www.getty.edu/research/conducting_research/vocabularies/ulan/(Accessed 1 December 2009).
- WorldWideWebSize. http://www.worldwidewebsize.com/ (Accessed 1 December 2009).
- WWAR. http://wwar.com/ (Accessed 1 December 2009).