April 11-14, 2007
San Francisco, California

Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques

Jacco van Ossenbruggen, Alia Amin, Lynda Hardman, and Michiel Hildebrand, CWI Amsterdam; Mark van Assem, Borys Omelayenko, Guus Schreiber, and Anna Tordai, Vrije Universiteit, Amsterdam; Victor de Boer, Bob Wielinga, and Jan Wielemaker, Universiteit van Amsterdam; Marco de Niet and Jos Taekema, Digital Heritage Netherlands DEN, The Hague; Marie-France van Orsouw and Annemiek Teesing, Dutch Institute for Cultural Heritage ICN, Amsterdam, The Netherlands

Abstract

This paper describes ongoing work of a project aimed at exploiting Semantic Web techniques to support searching and annotating large cross-institutional digital-heritage collections. The project demonstrator contains multiple collections and multiple vocabularies. The architecture is fully based on Web standards. We show novel search and presentation techniques which make use of interoperability between the collections and between the vocabularies.

Keywords: Semantic Web, multiple thesauri, semantic search, search paradigms, large virtual collections, scalability

1. Introduction

The main objective of the MultimediaN E-Culture project is to demonstrate how novel Semantic Web and presentation technologies can be deployed to provide better indexing and search support within large virtual collections of cultural-heritage resources. The architecture is fully based on open Web standards, in particular XML, SVG, RDF/OWL and SPARQL. One basic hypothesis underlying this work is that the use of explicit background knowledge in the form of ontologies/vocabularies/thesauri is in particular useful in information retrieval in knowledge-rich domains such as cultural heritage. This paper gives some details about the internals of the demonstrator. The on-line version of the demonstrator can be found at: http://e-culture.multimedian.nl/demo/search.

Readers are encouraged to first take a look at the demonstrator before reading on. As a teaser we have included a short description of basic search facilities in the next section. We suggest you consult the tutorial (linked from the on-line demo page) which provides a sample walk-through of the search functionality. The current version of the demonstrator runs under Firefox 2.0. It runs also under other browsers, but users may experience some problems with the SVG support. We expect these SVG problems to be solved in all main browsers in the near future. As a project we are committed to Web standards (such as SVG) and are not willing to digress to (and spend time on) special-purpose solutions.

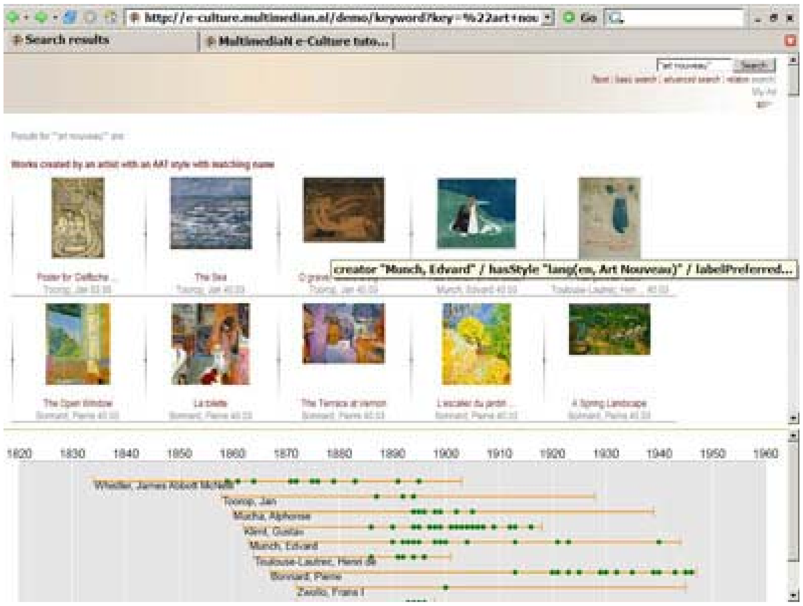
Please note that this is a product of an ongoing project. Visitors should expect the demonstrator to change. The project has a duration of 4 years and is at the time of writing 18 months underway. We are incorporating more collections and vocabularies and are also extending the annotation, search, and presentation functionalities.

This demonstrator won the Semantic Web Challenge (http://challenge.semanticweb.org) at last year’s International Semantic Web Conference, held in November 2006 in Athens, Georgia. This paper is a revised version of the description of this Challenge contribution.

2. A Peek at the Demonstrator

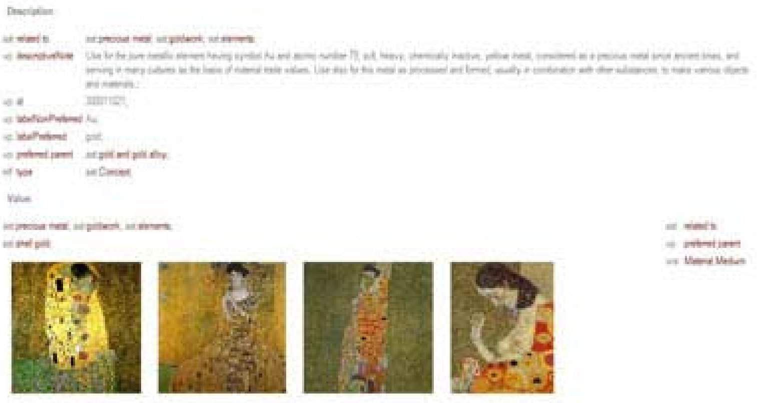
Figure 1 shows a query for “Art Nouveau”. This query will retrieve images that are related to Art Nouveau in some way. The results shown in the figure are “created by an artist with a matching style". So, these images are paintings by artists who have painted in the Art-Nouveau style, but the style is not part of the metadata of the image itself. This may retrieve some paintings which are not really Art Nouveau, but it is a reasonable strategy if there are no (or only few) images directly annotated with Art Nouveau. We view the use of such indirect semantic links as a potential for semantic search (for more details on path search in the demonstrator, see Section @@todo-path-search).

The lower part of the figure shows a listing of painters who are known to have worked in the Art-Nouveau style. The time line indicates years in which they have created art works (you can click on them to get information).



creator "Munch, Edvard" / hasStyle "lang(en, Art Nouveau)" / labelPreferred...

Fig 1: Results of query for “Art Nouveau”



Description

aat: related to aat: precious metal, aat: goldwork, aat: elements

vp: descriptiveNote Use for the pure metallic element having symbol Au and atomic number 79, soft, heavy, chemically inactive, yellow metal, considered as a precious metal since ancient times, and serving in many cultures as the basis of material trade values. Use also for this metal as processed and formed, usually in combination with other substances, to make various objects and materials.

vp: id 300011021

vp: labelNonPreferred Au

vp: labelPreferred gold

vp: preferred parent aat: gold and gold alloy

rdf: type skos: Concept

Value

aat: precious metal, aat: goldwork, aat: elements — aat: related to

aat: shell gold — vp: preferred parent

vra: Material Medium

Fig 2: Information about the indexing term “Gold”, showing also images that are related to “Gold”

Images have annotations in which terms from various thesauri are used. Figure 2 shows the information a user gets when selecting such an indexing term, here “Gold” material from the Art & Architecture Thesaurus. We also show images that have been annotated with this indexing term (or semantically related terms).

Figure 3 shows a snapshot of the facet browser. Here we have opted to look for “Paris” as a “depicted place”, i.e. art works with Paris as subject. As we see, the demonstrator also finds artworks about Montparnasse and Montmartre, as the THN thesaurus indicates that these are part of Paris. We can use the

TGN hierarchy to navigate to higher or lower parts in the hierarchy, e.g. to France. This browser would then show all art works with (some part of) France as subject. We can also cross-index it with other facets, e.g. all works of an expressionist style showing Paris.

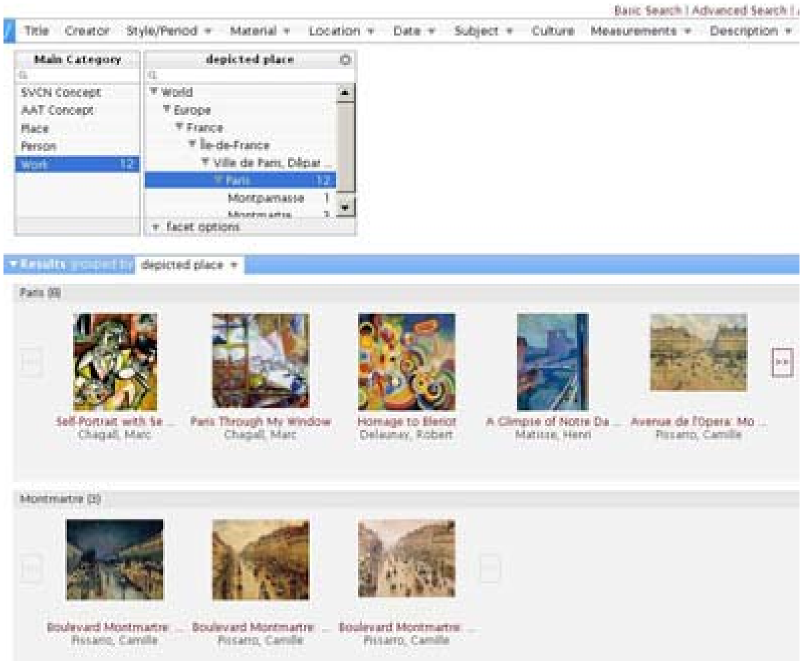


Fig 3: Facet browser snapshot, showing art works depicting “Paris”

These are some basic search and presentation functions. The reader is referred to the on-line demo for information about more search options, such as time-based ones. We also have an experimental search function for finding the semantic relations between two URIs, e.g. for posing the question ”How are Van Gogh and Gauguin related?" In fact, this leads to a whole avenue of new search possibilities and related issues with respect to which semantic paths are most relevant, and which we hope to explore in more detail in the coming years.

3. Technical Architecture

The foundation of the demo is formed by SWI-Prolog and its (Semantic) Web libraries (for detailed information, see Wielemaker 2003, 2005). SPARQL-based access is a recent feature. The technical architecture of the demonstrator is depicted in Figure 4. The “Application Logic” module defines searching and clustering algorithms using Prolog as query language, returning the results as Prolog Herbrand terms. The “Presentation Generation” module generates Web documents from the raw answers represented as Herbrand terms.

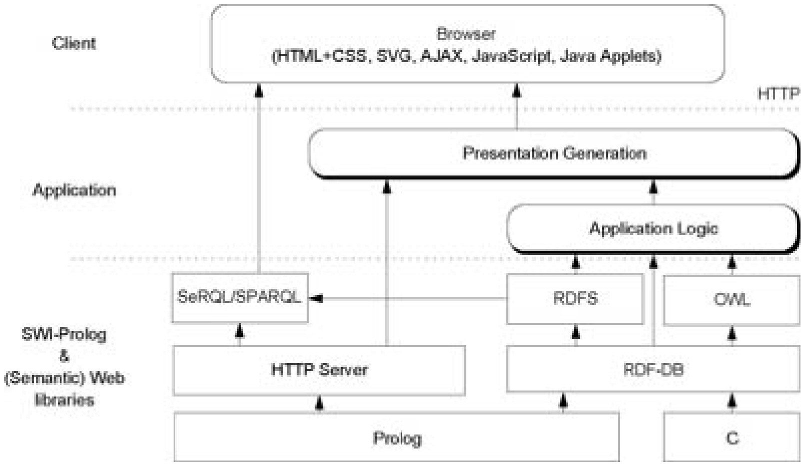


Fig 4: Technical architecture of the demonstrator (PDF)

From the user perspective, the architecture provides

1. annotation facilities for Web resources representing images, and
2. search and presentation/visualization facilities for finding images.

4. Vocabularies and Collection Data

4.1 Vocabularies

Currently, the demonstrator hosts four thesauri, namely the three Getty vocabularies (http://www.getty.edu/research/conducting_research/vocabularies/), i.e., the Art & Architecture Thesaurus (AAT), the Union List of Artists Names (ULAN) and the Thesaurus of Geographical Names (TGN), as well as the lexical resource WordNet, version 2.0. The Getty thesauri were converted from their original XML format into an RDF/OWL representation using the conversion methods principles as formulated by van Assem *et al.* (2004). The RDF/OWL version of the data models is available on-line (see http://e-culture.multimedian.nl/resources/). The Getty thesauri are licensed. The project has acquired licenses for the thesauri. People using the demonstrator do not have access to the full thesauri sources, but can use them to annotate and/or search the collections.

The RDF/OWL conversion of WordNet is documented in a publication of the W3C Semantic Web Best Practices and Deployment Working Group (van Assem, 2006). It is an instructive example of the issues involved in this conversion process, in particular the recipes for publishing RDF vocabularies (Miles et al., 2006).

The architecture is independent of the particular thesauri being used. We are currently in the process of adding the Dutch version of AAT, amongst others, to support a multi-lingual interface. Integration of other (multi-lingual) thesauri is planned.

Using multiple vocabularies is a baseline principle of our approach. It also raises the issue of alignment between the vocabularies. Basically, semantic interoperability will increase when semantic links between vocabularies are added. Within the Getty vocabularies, one set of links is systematically maintained: places in ULAN (e.g., place of birth of an artist) refer to terms in TGN. Within the project we are adding additional sets of links. One example is links between art styles in AAT (e.g. "Impressionism") and artists in ULAN (e.g., “Monet”). De Boer (2006) has worked on deriving these semi-automatically from texts on art history.

Finding semantic links between thesauri is part of the overall research efforts in ontology alignment: see, for example, the Ontology Alignment Evaluation Initiative (http://oaei.ontologymatching.org/). This is still a hard topic on its own, and the project intends to rely mainly on specialized efforts in this area to provide methods and techniques for semantic interoperability of vocabularies. On such efforts in the Dutch context is the STITCH project (http://www.cs.vu.nl/stitch).

4.2 Annotation Template

For annotation and search purposes, the tool provides the user with a description template derived from the VRA 3.0 Core Categories (Visual Research Association, 2000). The VRA template is defined as a specialization of the Dublin Core set of metadata elements, tailored to the needs of art images. The VRA Core Categories follow the “dumb-down” principle; i.e., a tool can interpret the VRA data elements as Dublin Core data elements. An unofficial OWL specification of the VRA elements, including links to Dublin Core, can be found at http://e-culture.multimedian.nl/resources/ .

4.3 Collection Data And Metadata

In principle, every Web resource with a URL can be included and annotated in the virtual collection of our demonstrator. As a test set of data we have included three Web collections:

1. The Artchive collection (http://www.artchive.com/) contains around 4,000 images of paintings, mainly from the 19th and 20th century.
2. The ARIA collection (http://rijksmuseum.nl/aria/) of the Rijksmuseum in Amsterdam contains images of some 750 master pieces.

3. The RMV collection (http://www.rmv.nl) of the “Rijksmuseum voor Volkenkunde” (State Museum for Ethnology) in Leiden describes about 80,000 images of ethnographic objects that belong to various cultures worldwide.

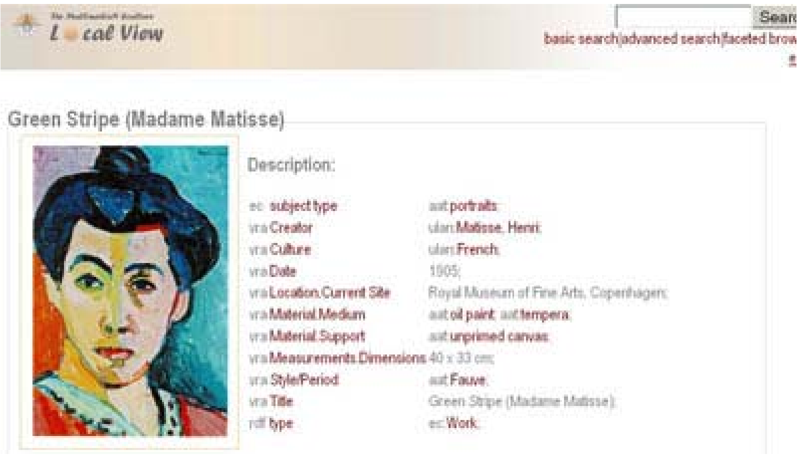
For the Artchive items, we have used a parsing technique to transform the existing textual annotation into a semantic annotation; i.e. matching strings from the text to concepts from the various thesauri.

The metadata that accompanies the Artchive collection consists of a short file holding textual values for title, creator, dimensions, material, year of creation, location and comments. Unfortunately the descriptor name is not specified with the value, and not all descriptions have the same values in the same order. We used a grammar to parse and canonize the date of creation and dimension fields. Author and material are matched to ULAN and AAT using a syntactic distance measure and selecting the best match.

For example, a famous painting of Matisse depicting his wife has the following textual annotation in Artchive:

> MATISSE, Henri
> Green Stripe (Madame Matisse)
> 1905
> Oil and tempera on canvas
> 15 7/8 x 12 7/8 in.
> Royal Museum of Fine Arts, Copenhagen

The resulting semantic annotation is shown in Figure 5. The “Style/period” property is not included in the original annotation text, but can be derived from the artist-style links present in the background knowledge.



Green Stripe (Madame Matisse)

Description:

ec: subject type	aat:portraits;
vra:Creator	ulan:Matisse, Henri;
vra:Culture	ulan:French;
vra:Date	1905;
vra:Location.Current Site	Royal Museum of Fine Arts, Copenhagen;
vra:Material.Medium	aat:oil paint; aat:tempera;
vra:Material.Support	aat:unprimed canvas;
vra:Measurements.Dimensions	40 x 33 cm;
vra:Style/Period	aat:Fauve;
vra:Title	Green Stripe (Madame Matisse);
rdf:type	ec:Work;

Fig 5: Metadata about the sample Matisse painting. The VRA properties are shown on the left; their values on the right. The values are prepended with the name space of the vocabulary that provides the term. For example, “ulan:Matisse, Henri;” refers to the ULAN entry for Henri Matisse.

For the other collections, we used similar strategies for enriching the original metadata with semantic categories. Adding a collection thus involves some information-extraction work on the metadata. In addition, the demonstrator supplies a manual-annotation interface which can be used to annotate any image on the Web.

5. Distributed Vs. Centralized Collection Data

The architecture is constructed to support multiple distributed image collections. Data (i.e. images) must have an external URI (we keep local copies, but that's only for caching). Ideally, we would like to get the original metadata also from external sources using standard protocols such as OAI (http://www.openarchives.org/). In practice however, we encountered several problems with the quality of metadata retrieved via OAI, so for the moment we still depend on the local copy of the original metadata.

Metadata extensions are also stored locally. In the future we hope to feed these back to the collection owners.

Vocabularies form a separate problem. The Getty vocabularies are licensed, so we cannot publish the full vocabulary as is. However, the information in the Getty vocabularies is freely accessible through the Getty online facilities; see, for example, http://www.getty.edu/research/conducting_research/vocabularies/aat/ for access to the AAT.

We hope that these vocabularies will become publicly available. In the meantime, our demonstrator allows you to browse the vocabularies as a semantic structure and search for images semantically related to a vocabulary item (see Figure 2 for an example of the concept “Gold” from AAT). An RDF/OWL version of WordNet has recently been published (see above). We will move within the next months to this version (the same version as we are now using but with a different base URI).

6. Keyword Search With Semantic Clustering

One of the goals of the demonstrator is to provide users with a familiar and simple keyword search, but still allow them to benefit from all background knowledge from the underlying thesauri and taxonomies. The underlying search algorithm consists of several steps that can be summarized as follows.

1. First, it checks all RDF literals in the repository for matches on the given keyword.
2. Second, from each match, it traverses the RDF graph until a resource of interest is found: we refer to this as a "target resource".
3. Finally, based on the paths from the matching literals to their target resources, the results are clustered.

To improve performance in finding the RDF literals that form the starting points, the RDF database maintains a btree index of words appearing in literals to the full literal, as well as a Porter-stem and metaphone (sounds-like) index to words. Based on these indexes, the set of literals can be searched efficiently on any logical combination of word, prefix, by-stem and by-sound matches (see http://www.swi-prolog.org/packages/semweb.html#sec:3.8).

In the second step, which resources are considered of interest is currently determined by their type. The default settings return only resources of type artwork (“vra:Work”), but this can be overridden by the user. To avoid a combinatorial explosion of the search space, a number of measures had to be taken. Graph traversal is done in one direction only: always from the object in the triple to the corresponding subject. Only for properties with an explicit owl:inverseOf relation is the graph also traversed in the other direction. While this theoretically allows the algorithm to miss out many relevant results, in practice we found that this is hardly an issue. In addition to the direction, the search space is kept under control by setting a threshold. Starting with the score of the literal match, this score is multiplied by the weight assigned to the property being traversed (all properties have been assigned a (default) weight between 0 and 1), and the search stops when the score falls under the given threshold. This approach not only improves the efficiency of the search, but also allows filtering out results with paths that are too long (which tend to be semantically so far apart that users do not consider them relevant anymore). By setting the weights to non-default values, the search can also be fine tuned to a particular application domain.

In the final step, all results are clustered based on the path between the matching literal and the target result. When the paths are considered on the instance level, this leads to many different clusters with similar content. We found that clustering the paths on the schema level provides more meaningful results. For example, searching on keyword "fauve" matches works from Fauve painters Matisse and Derain. On the instance level, this results in different paths:

dc:creator -> ulan:Derain -> glink:hasStyle -> aat:fauve -> rdfs:label -> "Fauve"

dc:creator -> ulan:Matisse -> glink:hasStyle -> aat:fauve -> rdfs:label -> "Fauve"

while on the schema level, this becomes a single path:

dc:creator -> ulan:Person -> glink:hasStyle -> aat:Concept -> rdfs:label -> "Fauve"

The paths are translated to English headers that mark the start of each cluster, and this already gives users an indication of why the results match their keyword. The path given above results in the cluster title, “Works created by an artist with matching AAT style". To explain the exact semantic relation between the result and the keyword searched on, the instance level path is displayed when hovering over a resulting image.

7. Vocabulary And Metadata Statistics

Table 1 shows the number of triples that are part of the vocabularies and metadata currently being used by the demonstrator. The table has three parts:

1. the schemas (e.g. the RDF/OWL schema for WordNet defining notions such as “SynSet”),
2. the vocabulary entries and their relationships, and
3. the collection metadata.

In total, these constitute a triple set of roughly 9,000,000 triples. We plan to extend this continuously as more collections (and corresponding vocabularies) are being added.

Document	Sources	#triples
Schemas		
RDFS./OWL	2	358
Annotation	6	769
Vocabularies	8	1,225
Collections	1	29,889
	Vocabularies	
TGN	4	425,517
ULAN	1	1,896,936
AAT	1	249,162
WordNet	18	2,579,206
	Collections	
Artchive	4	74,414
Rijksmuseum	1	27,933
RVM	1	3,662,257

Table 1: Number of triples for the different sources of vocabularies and collection metadata

Acknowledgements

The E-Culture project is an application subproject within the context of the MultimediaN ("Multimedia Netherlands", http://www.multimedian.nl) project funded by the Dutch BSIK Programme.

References

de Boer, V., M. van Someren and B. Wielinga (2006). “Extracting instances of relations from web documents using redundancy”. In: *Proc. Third European Semantic Web Conference (ESWC'06)*, Budva, Montenegro. http://staff.science.uva.nl/~vdeboer/publications/eswc06paper.pdf.

Miles, A., T. Baker and R. Swick (2006). Best practice recipes for publishing RDF vocabularies. Working draft, W3C Semantic Web Best Practices and Deployment Working Group. http://www.w3.org/TR/2006/WD-swbp-vocab-pub-20060314/.

van Assem, M., M. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga (2004). “A method for converting thesauri to RDF/OWL”. In McLlraith, S.A., D. Plexousakis, and F. van Harmelen (Eds.) *Proc. Third Int. Semantic Web Conference ISWC 2004*. Hiroshima, Japan. Volume 3298 of LNCS., Berlin/Heidelberg, Springer Verlag, pp. 17-31

van Assem, M., A. Gamgemi and G. Schreiber (2006). “Conversion of WordNet to a standard RDF/OWL representation”. In: *Proc. LREC 2006*. http://www.cs.vu.nl/~guus/papers/Assem06a.pdf

Visual Resources Association Standards Committee (2000). *VRA Core Categories, Version 3.0.* Technical report, Visual Resources Association. http://www.vraweb.org/resources/datastandards/vracore3/index.html

Wielemaker, J., G. Schreiber, B.J. Wielinga (2003). “Prolog-based infrastructure for RDF: performance and scalability”. In Fensel, D., K. Sycara and J. Mylopoulos (Eds.) *The Semantic Web - Proceedings ISWC'03*.

Sanibel Island, Florida. Volume 2870 of Lecture Notes in Computer Science. Berlin/Heidelberg, Springer Verlag, pp. 644-658.

Wielemaker, J., G. Schreiber, B. Wielinga. “Using triples for implementation: the Triple20 ontology-manipulation tool”. In Gil, Y., E. Motta, R. Benjamins and M. Musen (Eds.) *The Semantic Web - ISWC 2005: 4th International Semantic Web Conference*. Galway, Ireland, November 6-10, 2005. Volume 3729 of Lecture Notes in Computer Science., Springer-Verlag, pp. 773-785.

Cite as:

Ossenbruggen, J., et al., Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques, in J. Trant and D. Bearman (eds.). *Museums and the Web 2007: Proceedings*, Toronto: Archives & Museum Informatics, published March 1, 2007 Consulted August 10, 2015. http://www.archimuse.com/mw2007/papers/ossenbruggen/ossenbruggen.html

Editorial Note

published: April 11, 2007
last updated:October 28, 2010 12:23 PM

Archives & Museum Informatics, 158 Lee Avenue, Toronto, Ontario, M4E 2P3 Canada
Telephone: +1 416 691 2516 | Fax: +1 416 352 6025 | E-mail: mw2007 @ archimuse.com

Copyright © 2006 - Archives & Museum Informatics - All rights reserved.