

OpenART: Open Metadata for Art Research at the Tate

by Julie Allinson

Museum Informatics: Something New, Something More

EDITOR'S SUMMARY

The short-term OpenART project was geared to exposing linked open data for the research dataset, “The London Artworld 1660-1735,” of the Tate Britain art gallery, enabling viewers to learn about activities in the art world of the times. Work for the project required decisions and choices about using open data, starting with considering use cases and identifying entities and relationships. The dataset was modeled and expressed in the Resource Description Framework, using an event – art sales – as the central component of the model. An OpenART ontology was developed to provide domain-specific descriptors and integrated with existing ontologies to provide semantic interoperability. The resulting event-based ontology supported reciprocal statements about entities revolving around an event, as well as references back to other types of information for greater detail. Data was transformed from spreadsheets to RDF documents, served to a data repository and provided with persistent URLs. The OpenART project demonstrated practical approaches to exposing linked open data while enabling sharing and creative reuse of the research data..

KEYWORDS

fine arts	ontologies
linked data	open source
metadata	information reuse
research datasets	

Julie Allinson is digital library manager at the University of York, United Kingdom. Her overall responsibility is in managing the development of a multimedia digital library for the university, based on Fedora Commons software. She has a particular interest in metadata standards and exposing content to the widest audience, particularly with open and linked data. She can be reached at julie.allinson@at.york.ac.uk.

OpenART (<http://yorkdl.wordpress.com/category/openart>) was a six-month project funded by the Joint Information Systems Committee (JISC) in the United Kingdom under their Infrastructure for Resource Discovery Program. A partnership among the University of York, Tate Britain and technical partners Acuity Unlimited, the project’s aim was to design and expose linked open data for the Tate research dataset, “The London Artworld 1660-1735.” The creation of the dataset was partially funded by the Arts and Humanities Research Council (AHRC) as part of a larger project called “Court, Country, City.” The London Artworld 1660-1735 (hereafter London Artworld) is an ongoing research effort, involving Tate and the University of York, to transcribe primary and secondary sources that trace the people, places and activities comprising the art world in London at that time.

Partner efforts around this dataset provided a chance to enhance the planned website for London Artworld (<http://artworld.york.ac.uk/>) with semantic and open data. Open data in its broadest sense means data that can be made publicly available under a license that permits and promotes wide use. In our context, it also means data exposed in a format that allows for other applications to explore and use the data. In general, OpenART offered partners the opportunity to explore what open data might offer our respective institutions. The Tate, for example, is very interested in how open data might be beneficial throughout its systems, from provenance tracking through curation to its institutional website.

This article offers an overview of this short project, some reflections on the approaches chosen and some thoughts about open data in the library, archive and museum sector. It aims to be practical, focusing on the work done and choices made, more than on the theory behind it. The article assumes some familiarity with open and linked data. It is written from the

ALLINSON, continued

perspective of someone new to the whole area of open data and as such should not be read as a definitive guide.

The Data

Any project of this kind must, of course, start with some data. Use cases for how that data might be explored and re-used are also helpful. For OpenART, given the London art world subject matter, one use case is to help answer research questions such as these: Who were the members of the Rose and Crown Club? How many paintings were imported annually into England? Where did artists have their studios in this period? The answers to these questions may not be directly expressed or readily apparent or available in library, archive or museum information offerings or websites.

The OpenART data is a set of inter-linked spreadsheets, which are still being developed. The data is complex, with different columns, worksheets and spreadsheets, in differing states of completion. Two factors would prove very useful in addressing the complexity. First, that Richard Stephens, a researcher at the Tate and creator of the dataset, was engaged in the OpenART project and could provide his expert view of the data and second, that Richard had introduced a local set of unique reference numbers (URNs) for the different entities that he saw in the data.

FIGURE 1. Sketch showing entities and a selection of relationships among them.

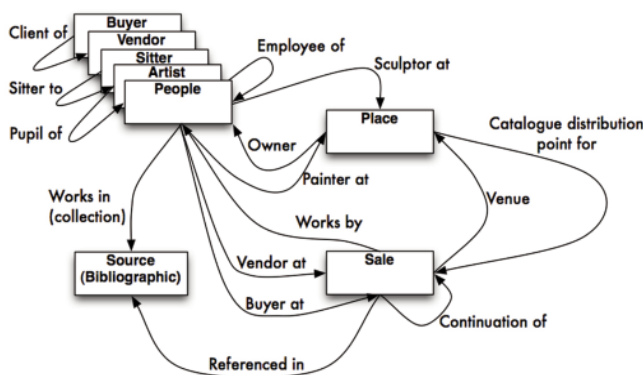


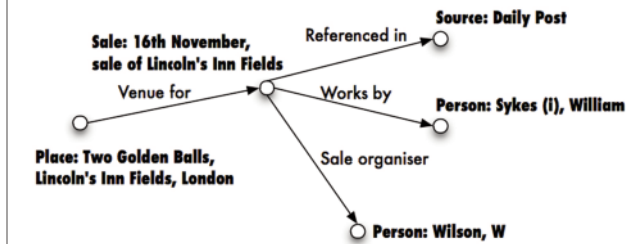
Figure 1 illustrates the types of entities and relationships in the data, as identified by Richard. The entities are People, Places, Sales and Sources – or, the kinds of art collecting that might, for

example, help answer questions like, how many paintings were imported annually into England? – along with other entities, such as Collection and Painting, which are

less fully-fleshed out at this stage in the original research data. The relationships among entities are expressed in the spreadsheets, with Richard’s unique URN used to show the links between sheets.

The first step for the University of York project team was to analyze the dataset and model it into something that could be expressed as open data, principally as Resource Description Framework (RDF). RDF is a W3C specification for modeling data. It is centered around the notion of triples (subject-predicate-object), where data becomes a graph of interconnected triples, as illustrated in Figure 2. For example, one triple illustrated in Figure 2 is Lincoln’s Inn Fields (subject) is the venue for (predicate) the sale of works by William Sykes (object).

FIGURE 2. Sketch showing a fragment of a graph for a sale.



The Model

Analyzing the data itself allowed us to explore the nature of the data and its characteristics. Although there are a variety of possible approaches, the one that was most resonant for us was the situational or event-driven model, given that much of the data was about tracing events at a certain place and at a specific time (London, 1660-1775). From our perspective, then, events were at the center of our data model, particularly art sales as events. Sales involve people and places and are expressed or evidenced by sources, such as newspapers, inventories and sale catalogues. While other events flow through the data – for instance, the act of purchasing (not always limited to sale), the act of advertising, artists working in particular locations during a particular period and sale catalogues being sold or distributed – given the project’s short timeframe, we narrowed the focus of our attention to art sales.

With general agreement on taking an event-driven approach, the next phase of work was to see what existing ontologies might help structure and model the data. Two existing ontologies looked useful for OpenART. One, the Linked Open Events Ontology (LODE – <http://linkedevents.org/ontology>) is a short ontology for describing events, which extends terms from other ontologies, including Dolce Ultra Lite (see below) and CIDOC-CRM, a conceptual reference model for cultural heritage information (www.cidoc-crm.org). LODE was also used by the LOCAH project (<http://data.archiveshub.ac.uk/>), another JISC-funded activity, which has rather blazed a trail for museums and archives wishing to expose open data. The other ontology of interest was Dolce Ultra Lite (DUL – http://ontology.designpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite), a lightweight upper ontology that describes general concepts across all knowledge domains.

Together, LODE and DUL could provide a general framework for describing the OpenART dataset, but they did not adequately express the domain specificity required of art research. For this reason, we decided to create our own ontology and offer more specific sub-classes for terms in LODE and DUL, thereby benefiting from the semantic interoperability that these existing ontologies provide, alongside the ability to provide rich, domain-specific descriptions.

For a short project, embarking on creating a new ontology was somewhat ambitious, and we were only able to take this approach because of the expertise within the University of York team. Alternative approaches such as using an existing ontology with no extension (quick and easy, but lacking domain-specificity) or mixing and matching terms from existing schemes coupled with local terms (flexible, but not fully reasoned) are also viable options.

The Ontology

The OpenART ontology was created by project partner Martin Dow at Acuity Unlimited. It is broken down into a number of ontologies, with one umbrella ontology bringing them together. Having separate ontologies should make the parts more readily understandable and reusable. The ontologies contain specific OpenART terms, their domains/ranges, superclass relationships and definitions. There are classes (that is, the entities), object

properties (for example, relationships between object classes) and data properties such as strings and dates. The project's ontologies can be viewed and downloaded from the OpenART Digital Library Ontologies page (<http://dlib.york.ac.uk/ontologies>).

While this article does not go into detail about ontology development – as it is rather a specialist activity – it is worth pointing out two useful tools: the Protégé ontology editor for ontology creation developed by Stanford University (<http://protege.stanford.edu>); and the Ontology Browser (<http://code.google.com/p/ontology-browser/>), an OWL ontology (see below) and RDF (Linked Open Data) browser. The Ontology Browser has been used to offer a more user-friendly presentation of the OpenART ontologies.

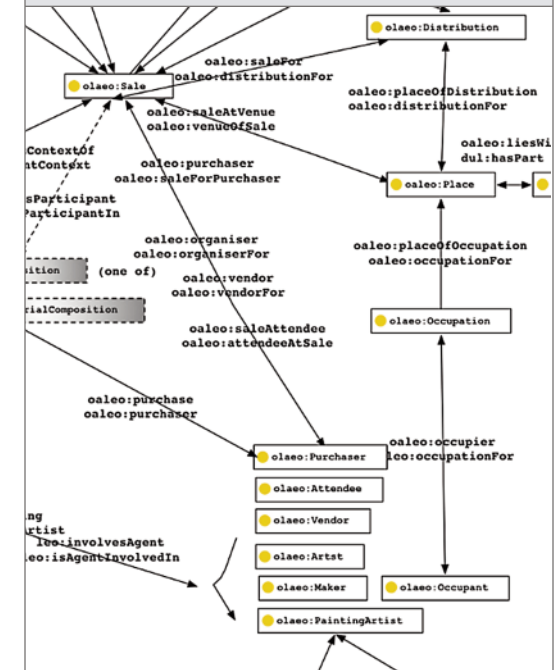
The fragment of the ontology shown in Figure 3 should provide a flavor of how the ontology works and its key aspects:

- 1) A situational or event-based ontology models one view on the data. Having one view means that rather than having, say, a Person with a *role* of Purchaser (and potentially multiple roles), there are specific *class names* for roles, such as the class Purchaser, for a Person involved in an event.

Purchaser P is purchaser at Sale S

We can infer from the superclasses (not shown in the fragment diagram; see the larger OpenArt ontology diagram [1]) that Purchaser

FIGURE 3. A fragment of the OpenART ontology, with entities and object properties; taken from a larger diagram.



is a Person and Sale is an Event, and thus our statement is both specific and broad, using only the one statement. This approach has the virtue of being simple, although it is not as comprehensive as one that could allow multiple contexts or views on the data.

- 2) We make reciprocal statements. In any instance data for a Sale; for example, our ontology allows us to make the reverse statement regarding the Purchaser.

Sale S is sale for purchaser Purchaser P
Purchaser P is purchaser at Sale S

- 3) A relationship between entities involves an event. Taking Distribution as an example, the place of distribution of a sale catalogue is centered on the act of Distribution, an event. This creates a set of relationships to convey the event of distribution and all of its participants (which can include objects, as well as persons).

Place P is the distribution point for Distribution D
Distribution D is the distribution event for Sale S
Catalogue C is participant in Distribution D

Note: C is not shown in the fragment diagram; see larger ontology diagram [1]

Creating sets of relationships among entities allows events to be described with greater specificity and, by reference to the superclasses, one can also understand the data in a broader context, in this case as some piece of information (Catalogue) being involved in some event (Distribution) happening at some place (Place).

There is much more to the ontology than described above. Other important aspects include describing source material using information object classes to describe the information (for example, the information in the catalogue) and using information realization objects for the physical realization of the information (for example, the actual printed catalogue). This distinction is important when using DUL and allows for a separation between what is known and what exists in physical form. The ontology also allows for granularity in describing parts of a source, for example the sale catalogue and its lot descriptions, and also for describing the paintings and artworks being sold.

Ours is an experimental approach, and in order to really test whether it is the right approach we need others to validate it by testing the ontology with their own data.

Making Data

Now that we had data (some spreadsheets) and a means of expressing that data (the ontology), our next step was to create open data and then to serve it up on the web. The ontology is written in a language called OWL, the Ontology Web Language, and one approach could have been to create instance data using the ontology, in OWL format. For our project, though, we wanted to generate RDF documents, one for each primary entity, the most widespread means of sharing open and linked data on the web. RDF is more widely consumed on the web and is a little simpler to create. The RDF document contains all of the triples about a given entity and some information about the document itself.

To get from the spreadsheets to RDF documents, we used a couple of tools. Google Refine (<http://code.google.com/p/google-refine>) proved an excellent tool for manipulating large spreadsheets and the RDF plug-in (<http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension>) allowed us to map data in the spreadsheets against our ontologies and to generate sample RDF data. In addition to this process, the project developer used java scripting to automate the processing after Refine had helped outline the structure, as well as to do some data clean-up.

For storing and serving up the data on the web, with such a short project, we opted to use the University of York's current application suite. In particular, we decided to store the generated RDF documents in our Fedora Commons repository, which acts as an enhanced filestore, and submit these documents to the external semantic indexing service SINDICE (<http://sindice.com>) to make them findable alongside other data, such as Wikipedia's open dataset (known as *dbpedia*).

We minted persistent URLs for our entities and their RDF data documents in the following pattern "dlib.york.ac.uk/id/entity/ref" where *entity* is one of Person, Place, Sale, Source, and *ref* is the numerical part of the persistent identifier from our Fedora repository; for example, 1234 represents the

ALLINSON, continued

persistent identifier york:1234. In this way, Fedora has a role in managing the identifiers and ensuring their uniqueness. URL rewriting and content negotiation have been implemented to match the identifier with the content's location in the University of York's digital library. These mechanisms ensure that when users access the URL they are served the correct resource and the correct document format they requested such as HTM or RDF/XML. Internally, Fedora Commons and our YODL digital library front-end application handle serving up the content to the web (<http://dlib.york.ac.uk>). The HTML view of the data is in actual fact a redirect to a different web application containing a human-readable view of the data. The YODL site, developed for the larger AHRC project, is in the process of having RDF/A embedded into its pages to offer another mechanism for machine crawling of the data.

The data itself exists in RDF documents, as mentioned above. To simplify things, we only created persistent identifiers and documents for each primary entity, those mentioned at the very beginning of this piece: Person, Place, Sale, and Source – the things we think art researchers are looking for, around the topic of the London Artworld. Later, we may add identifiers and documents for artworks, but those data aren't ready yet. Other entities in the data such as an occupation event like Appraisal or Auction are described using blank nodes, an RDF construct for describing entities that are not identified externally but are vital to produce valid data.

A simple sample RDF document for a Place can be seen below. This sample contains a direct relationship to a Sale venue and a blank node for Occupation, which links Place to Person via an Occupation node to express the notion of an Artist that worked in this Place. We decided that the Occupation entity did not need an identifier, although another implementer of the ontology might justifiably mint identifiers for every entity. The document format used here is TURTLE, an RDF serialization that is much easier to understand by eye. Other sample RDF documents can be viewed on the data page of the OpenART Digital Library (<http://dlib.york.ac.uk/data/>).

```
<http://dlib.york.ac.uk/id/place/3_0175>
  mapping:hasResearcherID "3.0175"^^<xsd:string>;
  rdfs:label "The house of John Stone in St Martin's Lane";
  vocupper:hasPlaceName "The house of John Stone in St Martin's Lane";
```

```
vocupper:hasBuildingName "house of John Stone";
vocupper:hasStreetName "St Martin's Lane";
vocupper:hasCounty "Metropolitan London";
vocupper:hasCountry "England";
vocupper:hasPlaceDescription "Stone's house was 'in the upper end
  of St.Martins Lane, next to St.Giles's Fields.'";
oactxt:venueOfSale <http://dlib.york.ac.uk/id/sale/2_0389>;
oactxt:placeOfOccupation [
  oactxt:occupationFor <http://dlib.york.ac.uk/id/person/1_00001>;
  rdf:type oactxt:Occupation, owl:NamedIndividual
];
vocupper:liesWithin [
  rdf:type model:Place, vocupper:County, owl:NamedIndividual
];
vocupper:liesWithin [
  owl:sameas <http://www.geonames.org/6269131/>;
  rdf:type model:Place, vocupper:Country, owl:NamedIndividual
].
```

Conclusions from OpenART

OpenART was a tiny project but we hope that it shows one approach to describing rich, specific datasets in a way that can be used in both broad and specific contexts. The narrow focus of our project – on art sales – could easily extend to events throughout the lifecycle of an artwork or to the lifetime of artist or art venue and beyond. Hooking up with existing datasets and authority sources could enhance the London Artworld data and make it more linked and linkable. One such example is using the unique identifiers from the Getty's Union List of Artist Names (www.getty.edu/research/tools/vocabularies/ulan/), published via the Virtual International Authority File initiative (VIAF – <http://viaf.org/>).

Our project needs wider partnerships and validation to see whether our experimental approach in using an event-based ontology has value for the community. We are at the very early stages of our investigation, only now at

ALLINSON, continued

the point of releasing our data on the OpenArt website. Although funding is now at an end, the work continues in enhancing the data, testing with tools and indexing services and looking out for other data-linking opportunities.

If there are lessons to pass on, they would be the following:

- 1) Developing an ontology takes time and expertise. If you don't have these resources, choose an approach that has been used before; add to it, go with it and accept that it may not be a perfect fit.
- 2) Working with moving targets is hard. If your data is changing, work with the data creators to understand the dataset and what is being added and changed, and make sure they understand the impact of changes on the exposure of data.
- 3) Learn as much as you can about linked data. Read a book like *Linked Data: Evolving the Web into a Global Data Space* [2] and don't be scared off. At its simplest, linked data is really very simple.
- 4) There are plenty of tools out there for creating and exposing linked data, some good for local installations and some for externally hosted

solutions. Find the solution(s) that best fits your local technical expertise and capacity (or lack of it).

Final Thoughts on Why?

So, why open data, why linked data? What's the business case for putting effort into mapping and making data like this?

Linked data is a growing area, and the full benefits of today's efforts to make open data may not immediately be clear. In the longer-term though, a greater pool of data from trusted sources should mean reduced duplication and greater sharing among institutions, which in turn means greater efficiencies and wider exposure. And while OpenART alone did not offer the Tate direct solutions to applying open data throughout its many systems, the project opened up a valuable discussion of why and how. By taking the leap of putting open data out on the web, such as has been done with the London Artworld, we are opening it up to other users and institutions so that they may mix it into their data, create visualizations and potentially create interesting and dynamic new resources that might never have been imagined or funded without open data. ■

Resources Mentioned in the Article

[1] Diagram of the OpenART ontology: <http://dlib.york.ac.uk/yodl/app/image/detail?id=york:28365>

[2] Heath, T., & Bizer, C. (2011) *Linked Data: Evolving the Web into a Global Data Space*. (Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1). Morgan & Claypool.