

# How Linked Open Data Can Help in Locating Stolen or Looted Cultural Property

Eleanor E. Fink<sup>1</sup>, Pedro Szekely<sup>2</sup>, and Craig A. Knoblock<sup>2</sup>

<sup>1</sup> 2360 North Vernon Street, Arlington, Virginia 22207  
eleanorfink@earthlink.net

<sup>2</sup> University of Southern California, Information Sciences Institute  
4676 Admiralty Way, Marina del Rey, CA 90292  
{pszekely,knoblock}@isi.edu

**Abstract.** Looting and theft of cultural property has been a problem for decades. While there are no exact figures, some agencies suggest it is a criminal industry grossing in the billions annually. Documentation is an essential and key component to finding lost or stolen cultural property and in establishing ownership in a court of law. However, the data on cultural heritage is locked up in data silos making it exceptionally difficult to search, locate, and obtain reliable documentation. Through an advancement of the Semantic Web, called Linked Open Data (LOD), walls can disappear and the potential for a global database on cultural heritage becomes possible. We will introduce and demonstrate how LOD is produced and point to new tools such as Karma that can handle conversion of large quantities of cultural heritage data to LOD. With LOD and a tool like Karma we can establish bridges across repositories of information and simplify access to cultural heritage information that in the long term could help protect cultural property from looting and theft.

**Keywords:** Semantic Web, Linked Open Data, RDF, ontology, CIDOC CRM, art theft, looting, cultural property, global cultural heritage database, information retrieval

## 1 Introduction

Looting and theft of cultural property is a worldwide problem often resulting in the destruction or loss of a piece of the history of mankind. Law enforcement agencies such as Interpol and Europol state that without documentation it is almost impossible to recover stolen or looted cultural property. A good example is the case of the Kanakaria mosaics that were stolen from a Greek Orthodox Church in the Turkish-occupied area of Cyprus and later turned up in the possession of a US art dealer in Indiana. Unfortunately, Cyprus did not have documentation that could be introduced in a court of law to prove ownership of the mosaics. However, attorney Thomas R. Kline, who represented Cyprus in the restitution case, was able to locate and introduce a publication by Dumbarton Oaks that contained an article with photographs illustrating the mosaics

in situ in the Church of the Kanakaria before they were looted. As a result of the photographs and documents in the publication, the judge ruled in favor of Cyprus.

It is ironic that on the one hand there is a wealth of information about stolen and looted cultural property in publications and on the Web, but much of that data is locked up in data silos making it difficult for law enforcement and legal experts to locate it. Adding to the irony is that the technology exists to break down the silos making it possible to simultaneously search and browse information from several websites at a time. The stumbling block is (1) lack of awareness of technology advances such as Linked Open Data (LOD) [1] and tools for mapping data to LOD, such as Karma, and (2) lack of policies that would lead to open access across art theft databases, museums, and agencies reporting stolen cultural property.

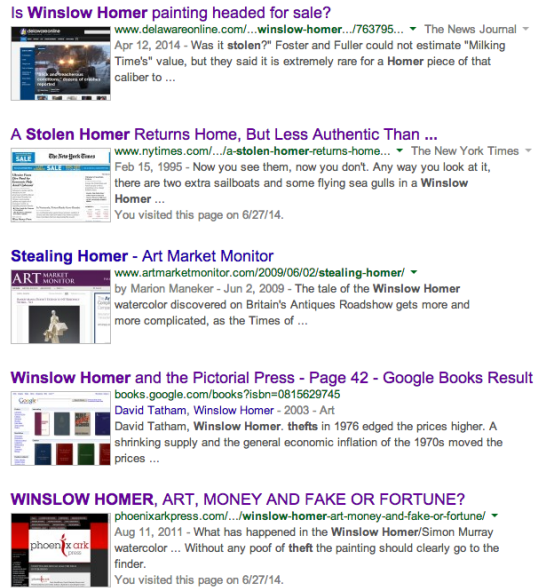
In this paper we present a vision of how Linked Open Data (LOD) can provide an integrated view of all of the relevant data, including the collection managements systems in museums, the hundreds of databases about stolen and looted art, auction databases, databases with provenance information, and on-line communities and blogs. LOD can be the platform to create a worldwide, tightly interconnected, searchable and analyzable “global database” of information to address the problem of finding documentation about stolen and looted art without necessarily knowing where to look. But to achieve this vision we need an appropriate tool such as Karma that has successfully demonstrated conversion of cultural heritage information to LOD.

## 2 Searching for Looted or Stolen Cultural Property

Let us consider a concrete example to contrast the current technology to find information about stolen art, and the technology enabled by LOD. Several paintings of American painter Winslow Homer have been reported stolen or missing and recovered in the last 40 years. We cannot go on the Web today and identify the stolen ones. Nor can we find a list of all known Winslow Homer paintings. A Google search for “winslow homer theft” provides “About 1,140,000 results”. Figure 1 shows five entries in the first page of search results. The search results are documents, not information, so we need to open and read the documents to determine whether they are relevant. In these five results we can identify two stolen Homers, “Off Gloucester Harbor” and “Children Under a Palm”. The process of assembling answers from Google search results is labor intensive and unreliable. The first 5 pages of results all contain the words “winslow homer” and “theft” in articles such as “Grooming Women for Leadership”, but no information about “Boy Reading” a stolen Homer listed in the FBI National Stolen Art File. The FBI file is not one stop shopping either given that it has no information about other stolen Homer artworks. Web sites such as <http://illicitculturalproperty.com/> and <http://obs-traffic.museum/> report no results when searching for “homer”. Finding a list of all Homer art-

works is equally daunting as it would require searching many Web sites, evaluating the provenance of the data, and manually assembling the list.

In contrast, LOD is a precisely searchable database containing information from all relevant data sources including databases, collection management systems, Web sites, blogs, news aggregators, etc. Instead of documents, LOD represents the information as entities (e.g., an entity for Winslow Homer, entities for each of his artworks, and entities for events such as ownership transfers and thefts). The entities are linked, so that for example, one can query for all Winslow Homer artworks or all theft events. The entities are also linked to the original sources so that one can verify provenance and find additional information. The seamless access and precise search capabilities of this “global database” would dramatically enhance law enforcement’s ability to find looted or stolen cultural property. But achieving the benefits of LOD and innovative concepts, such as a “global database” on cultural heritage, requires both social progress to open the many closed resources that exist about stolen and looted art and technology tools, such as Karma.



**Fig. 1.** A Google search for “winslow homer theft” retrieves documents that users must read to extract relevant information.

### 3 Linked Open Data

In computing, Linked Data describes a method of publishing and linking structured data so that it becomes more useful [1]. To achieve greater context and meaning in documents, pieces of information have to be tagged much like XML for publishing on the Web. In the case of LOD, a language called RDF [7] is used for tagging the published data. RDF breaks down knowledge into discrete pieces, with rules about the semantics, or meaning, of those pieces. Information is expressed as a list of statements in the form subject/predicate/object, known as triples. Each subject, predicate, and object can be represented with a Uniform Resource Identifier (URI) [8]. An ontology must be selected to play the key role of defining the meaning of the terms used in the statements. In essence, RDF along with an ontology insert context and meaning to a statement and the

URIs provide the unique identifiers and ability for all the pieces of the triples published as LOD to be searched and connected.

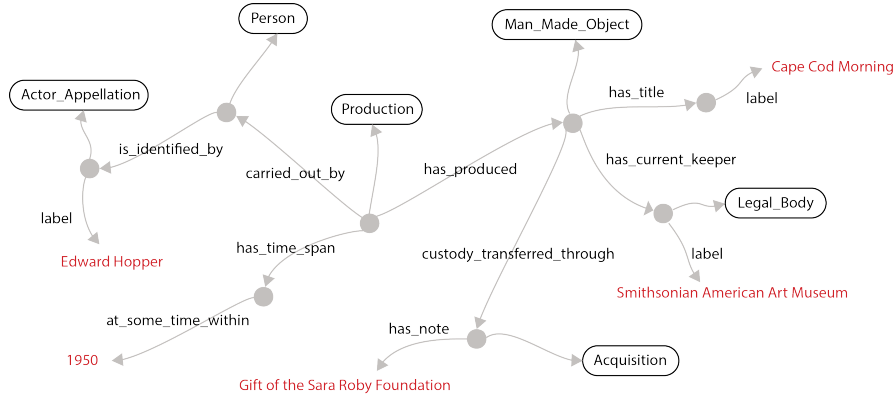
Since pieces of information are tagged to indicate their precise meaning, it becomes possible to search for information rather than documents. Until recently, information resources on the Web were limited to hyperlinks that create connections on a document level. Although these connections are useful by enabling one to click through to various resources, they do not express what type of connection there is between two pages, merely that there is one. For example, when using LOD, the relationship between artist and work of art is explicitly tagged, enabling accurate search results. A search for LOD-tagged information about the Venus de Milo will result in links to the ancient Greek marble sculpture in Paris at the Louvre, as opposed to tennis star Venus Williams or the planet.

The information about things in LOD format is represented in an information cloud. Today that cloud already contains billions of pieces of information about people, places, and things [1]. In respect to searching for a particular work of art, the key feature of LOD highlighted in this paper is that it allows for searching across institutions that use LOD to list and document cultural objects. The data silos that currently separate one institution or museum website from another disappear. Instead of silos, one can search several institutional documents and collections at a time. But knowledge of LOD within the cultural heritage domain is still nascent. One of the major challenges is how to efficiently convert data from every museum or cultural institution into LOD format to make the “global database” a reality.

## 4 Karma: A Tool for Publishing Linked Open Data

The cultural heritage community adopted the CIDOC Conceptual Reference Model [2] (CRM) ontology for describing information about cultural heritage. The CRM, an ISO standard since 2006, is designed to model “all information required for the scientific documentation of cultural heritage collections, with a view to enabling wide area information exchange and integration of heterogeneous sources.” The CRM is designed to support precise descriptions, and consequently has a large number of terms (82 classes and 263 properties). Mapping cultural heritage data to the CRM is difficult, so if we expect every cultural institution in the world to map their data to LOD, it is crucial to give them easy-to-use tools so that they can produce and maintain the mappings at low cost.

Figure 2 illustrates the richness of the CRM and the complexity of mapping data from a museum collection management system to the CRM. The figure shows a fragment of the CRM representation for Edward Hopper stating that Edward Hopper is a person and that he carried out the production of a man-made object whose title is Cape Cod Morning; this object was transferred to SAAM as a gift from the Sara Roby Foundation. The complete record for Edward Hopper is significantly larger, and gets assembled from data stored in multiple records in multiple tables in the collection management system. The LOD for



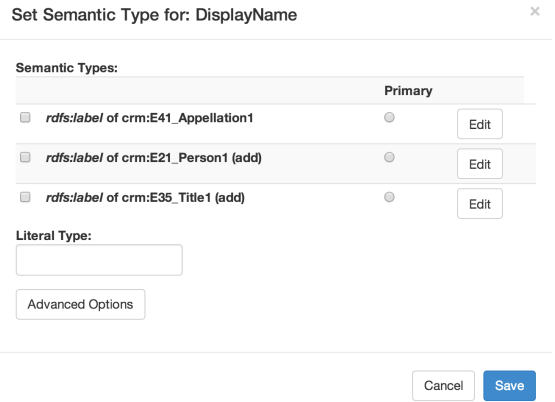
**Fig. 2.** LOD representation of “Edward Hopper produced Cape Cod Morning in 1950, kept at the Smithsonian American Art Museum, acquired as Gift of the Sara Roby Foundation”. The LOD data is represented using the CIDOC CRM ontology.

the complete SAAM collection (44,000 object records) consists of a network with over 3,000,000 edges [10].

The challenge when mapping data to CRM is that each column in each table in the collection management system must be mapped to the appropriate class and property in the CRM ontology. For example, the cell in the database containing the text *Edward Hopper* must be mapped to the label of an Actor\_Appellation. Then, each of the individual fragments for each cell must be connected together using the appropriate properties (represented as the labeled arrows connecting the grey circles).

The process is intellectually challenging because the resulting structures are elaborate, and the appropriate terms must be used to label the nodes and the arrows to accurately capture the meaning of the data in the collection management system. The process is also technically challenging because we need

to write executable specifications that a software program can use to automatically generate these structures for thousands or millions of objects, correctly handling the idiosyncrasies of every single object to produce the appropriate



**Fig. 3.** Karma suggests semantic types for attribute DisplayName learned from prior semantic type assignments in other datasets.

RDF for it. Furthermore, every cultural institution is different, requiring different data-to-RDF specifications even when they use the same collection management system. Often, the data needs to be cleaned before it can be mapped, adding significant complexity and cost to the process.

In previous work, we developed Karma [5, 10], a tool to enable data-savvy users (e.g., spreadsheet users) to define the data-to-ontology mapping specifications, shielding them from the complexities of the underlying technologies (SQL, SPARQL, graph patterns, XSLT, XPath, etc). Karma addresses this goal by automating significant parts of the process, by providing a visual interface where users see the Karma-proposed mappings and can adjust them if necessary, and by enabling users to work with example data rather than just schemas and ontologies.

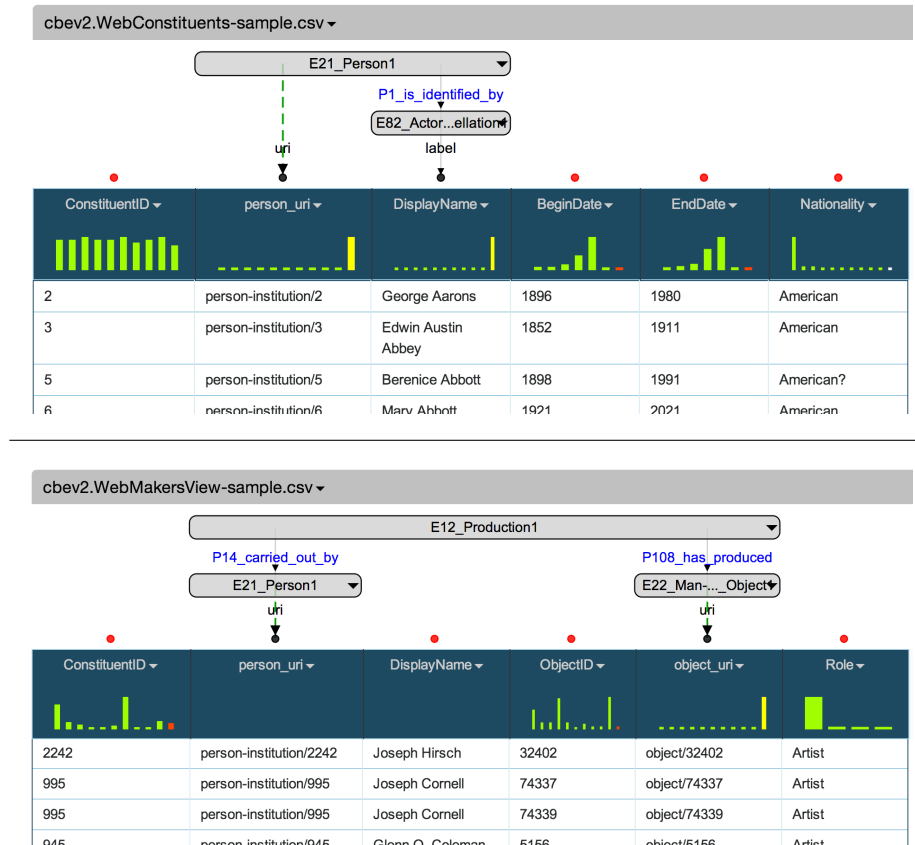


Fig. 4. Partial mapping of artists (top), and of objects (bottom).

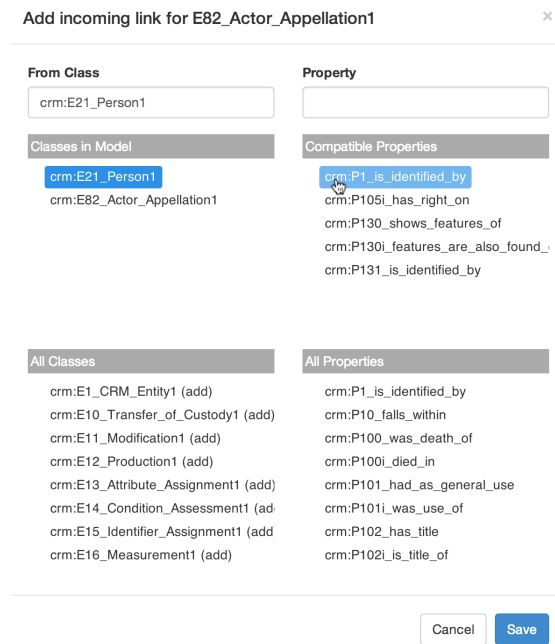
The Karma approach to map data to ontologies involves two interleaved steps: one, assignment of semantic types to data attributes and two, specification

of the relationships between the semantic types. A semantic type specifies the meaning of a single attribute relative to an ontology. For example, consider an attribute called `DisplayName` that stores the names of artists. In our approach, the semantic type for this attribute represents its meaning, for example, the label of a `E82_Actor.Appellation`, where `label` is a property in the ontology and `Person` is a class in the ontology. In general, a Semantic Type specifies the mapping of a single attribute to a property and corresponding class in an ontology. Karma uses a conditional random field (CRF) [6] model to learn the assignment of semantic types to columns of data from user-provided assignments. Karma uses the CRF model to automatically suggest semantic types for unassigned data columns. Figure 3 illustrates the capabilities of the learning component: Karma offers three suggestions for the semantic type for the `DisplayName` attribute. The first suggestion is correct, and the user can simply select it to define the semantic type for `DisplayName`. When the desired semantic type is not among the suggested types, users can browse the ontology to find the appropriate type. Karma automatically re-trains the CRF model after these manual assignments.

The relationships between semantic types are specified using paths of object properties. Given the ontologies and the assigned semantic types, Karma creates a graph that defines the space of all possible mappings between the data source and the ontologies. The nodes in this graph represent classes in the ontology, and the edges represent properties. Karma then computes the minimal tree that connects all the semantic types, as this tree corresponds to the most concise model that relates all the columns in a data source, and it is a good starting point for refining the model (Figure 4).

Sometimes, multiple minimal trees exist, or the correct interpretation of the data is defined by a non-minimal

tree. For these cases, Karma provides an easy-to-use GUI to let users select a desired relationship (an edge in the graph). Figure 5 shows the interface for adjusting the links between classes. In this example, the user is adjusting the link to the `E82_Actor.Appellation` class. When the user selects the source of the



**Fig. 5.** Karma interface for adjusting links offers suggestions of compatible properties between the source and the destination classes.

link (E21.Person), Karma suggests properties based on the definition of the ontology. In our example, the correct property (P1.is\_identified\_by) is in the list of suggestions. Karma also allows the user to browse all the properties in the ontology for cases when the suggestions do not contain the desired property.

After mapping the data to the ontology, Karma can be used in batch mode to generate the RDF data for the full contents of the data. The process is efficient, taking less than 5 minutes (on a laptop computer) to generate over 3,000,000 RDF triples for the 40,000 objects in the SAAM collection management system.

Once Karma generates the RDF data, the next step is to interlink it with other Linked Data in the LOD cloud. The interlinking process involves finding the URIs in the LOD cloud that refer to the same entities (e.g., finding the URIs for Winslow Homer) and asserting that they are equivalent. Many automatic interlinking tools exist [4,9], but even the best ones are not 100% accurate. The state of the art tools seldom are more than 95% accurate. For this reason, Karma offers a link curation tool that enables curators to verify the links. The tool records the provenance of the links and the human decisions before publishing the links in the LOD cloud. Figure 6 shows a screenshot of the verification tool. In this case the user is being asked to verify links from the SAAM dataset to DBpedia (the LOD version of Wikipedia) and the NYTimes Linked Data. Each record represents a proposed link. The top part shows data from the SAAM database, and the bottom part shows the data from DBpedia. The user can click on the records to read the full description and then can use the buttons on the right to verify the link and enter a comment. The system records the complete history of link verifications.






















 Person_5502  	1917	2009	Andrew Wyeth	Verified by Human	<input type="button" value="Match"/> <input type="button" value="Not Match"/> <input type="button" value="Unsure"/>
	1917	2009	Andrew Wyeth	 [2012-11-08 12:24:09] History	
 Person_1483  	1607	1656	Aniello Falcone	Exact match (0.99999704)	<input type="button" value="Match"/> <input type="button" value="Not Match"/> <input type="button" value="Unsure"/>
	1600	1665	Aniello Falcone	[2012-11-21 15:54:22] History	
 Person_3946  	1895	1978	Abraham Rattner	Verified by Human	<input type="button" value="Match"/> <input type="button" value="Not Match"/> <input type="button" value="Unsure"/>
	1893	1978	Abraham Rattner	 [2012-11-08 12:24:07] History	
 Person_70  	1905	1978	Arturo Pacheco 	Verified by Human	<input type="button" value="Match"/> <input type="button" value="Not Match"/> <input type="button" value="Unsure"/>
	1903	1978	Arturo Pacheco 	 [2012-11-08 12:24:10] History <input type="text" value="comment here"/> <input type="button" value="✓"/> <input type="button" value="✕"/>	
 Person_18387  	1948	----	Abelardo Morell	Verified by Human	<input type="button" value="Match"/> <input type="button" value="Not Match"/> <input type="button" value="Unsure"/>
	1948	----	Abelardo Morell	 [2012-11-08 12:24:06] History	

Fig. 6. Partial mapping of artists (top), and of objects (bottom).

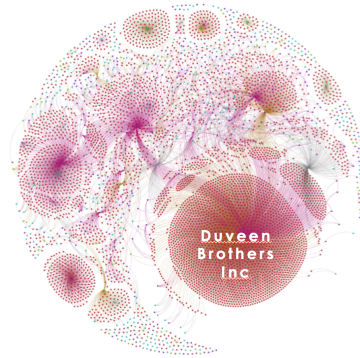


## 5 Towards the Global Cultural Heritage LOD

Our work so far focused on using Karma to map data from collection management systems to LOD, eliminating the silos across museums. Augmenting it with LOD data from law enforcement databases is the next natural step that can be accomplished using the same tools. Further enrichment would involve incorporation of existing provenance data, art auction repositories, and information extracted from both online forums and news stories.

Recent work in our group shows that technology is ready to tackle these further steps. In a demonstration project, we extracted provenance records from the Web pages of the National Gallery of Art, used provenance data in CSV files from the Getty Provenance Index Databases, converted all records to LOD, and interlinked it with artist information from Wikipedia. To show the benefit of LOD, we constructed visualizations of the LOD provenance networks that show the social network of art dealers in our demonstration dataset (Figure 7).

The tools to extract data from Web sites and text documents have matured significantly in recent years, making it possible to produce XML or JSON documents for Web sites that do not yet provide access via an API. Once the XML or JSON data has been extracted, our Karma tool can then be used to convert the data to LOD.



**Fig. 7.** Social network of art dealers in the provenance LOD.

## 6 Conclusion

There are no silver bullets to stop illicit trafficking of cultural property or guarantee recovery of stolen and looted items. But there are steps that could be taken to use innovations such as LOD that would make a difference. With the advancement of the Semantic Web and Linked Open Data, the idea of one stop searching or seamless access across silos of information is achievable if more agencies agreed to produce their data in LOD format.

If we are serious about trying to stem illicit trafficking and speed up recovery of missing works, there needs to be international cooperation in forging agreement on standards, a movement away from data silos, and international support of open access policies and use of Linked Open Data.

In order to fully explore and demonstrate how LOD can eliminate data silos, we formed the American Art Collaborative (AAC) [3], comprised of fourteen museums with American art collections from across the United States, namely Amon Carter Museum of American Art in Fort Worth, TX; the Archives of American Art, Smithsonian Institution, in Washington, D.C.; the Autry National

Center in Los Angeles, CA; the Colby College Museum of Art in Waterville, ME; Crystal Bridges Museum of American Art in Crystal Bridges, AR; the Dallas Museum of Art in Dallas, TX; the Thomas Gilcrease Museum in Tulsa, OK; the Indianapolis Museum of Art in Indianapolis, IN; the Metropolitan Museum of Art in New York, NY; the National Museum of Wildlife Art in Jackson Hole, WY; the National Portrait Gallery, Smithsonian Institution, in Washington, D.C.; the Princeton University Art Museum in Princeton, NJ; the Smithsonian American Art Museum in Washington, D.C.; and the Walters Art Museum in Baltimore, MD. We used Karma to map the data from several museums in the AAC. In addition to SAAM, we mapped the data from the Amon Carter and Crystal Bridges museums. We are currently working to map the data from the other museums and to verify the links between the datasets. In addition to opening up access as well as enabling searches across their combined collections, AAC participants strongly value being the authoritative source for LOD records about their collections as opposed to LOD records produced by third party aggregators. Providing an authoritative and reliable source adds value if the documents are to be used in a court of law to prove ownership.

## References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)* (2009), <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>
2. Doerr, M.: The cidoc conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag.* 24(3), 75–92 (Sep 2003), <http://dl.acm.org/citation.cfm?id=958671.958678>
3. Fink, E., Richey, S., Szekely, P.: (2013), <http://americanartcollaborative.org>
4. Isele, R., Bizer, C.: Learning expressive linkage rules using genetic programming. *Proceedings of the VLDB Endowment* 5(11), 1638–1649 (2012)
5. Knoblock, C.A., Szekely, P., Ambite, J.L., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyan, M., Mallick, P.: Semi-automatically mapping structured sources into the semantic web. In: *The Semantic Web: Research and Applications*, pp. 375–390. Springer (2012)
6. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
7. Lassila, O., Swick, R.R.: (1999), <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>
8. Masinter, L., Berners-Lee, T., Fielding, R.T.: Uniform resource identifier (uri): Generic syntax (2005), <http://tools.ietf.org/html/rfc3986>
9. Ngomo, A.C.N., Auer, S.: Limes: a time-efficient approach for large-scale link discovery on the web of data. In: *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*. pp. 2312–2317. AAAI Press (2011)
10. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the smithsonian american art museum to the linked data cloud. In: *The Semantic Web: Semantics and Big Data*, pp. 593–607. Springer (2013)