Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article

Marcia Lei Zeng

How to cite this article:

Zeng, Marcia Lei (2019). "Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article". *El profesional de la información*, v. 28, n. 1, e280103.

https://doi.org/10.3145/epi.2019.ene.03

Article received on December 13th, 2018
Approved on December 28th, 2018

Marcia Lei Zeng ✉
https://orcid.org/0000-0003-0151-5156

Kent State University, School of Information
1125 Risman Drive, 314 University Library
Kent, Ohio, OH 44242-0001, USA
mzeng@kent.edu

Abstract

With the rapid development of the digital humanities (DH) field, demands for historical and cultural heritage data have generated deep interest in the data provided by libraries, archives, and museums (LAMs). In order to enhance LAM data's quality and discoverability while enabling a self-sustaining ecosystem, "semantic enrichment" becomes a strategy increasingly used by LAMs during recent years. This article introduces a number of semantic enrichment methods and efforts that can be applied to LAM data at various levels, aiming to support deeper and wider exploration and use of LAM data in DH research. The real cases, research projects, experiments, and pilot studies shared in this article demonstrate endless potential for LAM data, whether they are structured, semi-structured, or unstructured, regardless of what types of original artifacts carry the data. Following their roadmaps would encourage more effective initiatives and strengthen this effort to maximize LAM data's discoverability, use- and reuse-ability, and their value in the mainstream of DH and Semantic Web.

Keywords

Semantic enrichment; Libraries, archives, and museums; LAMs; Digital humanities; DH; Smart data; Metadata; Structured data; Semi-structured data; Unstructured data; Knowledge discovery; Entity-centric modeling and information access; Data integration and interoperation; Literature review.

1. Introduction

The role of libraries, archives, and museums (LAMs) in supporting digital humanities (DH) research and education has been widely recognized in recent years. In DH research, the difficult part is not typically at the stages of data's cleaning, analysis, visualization, and synthesizing. The most challenging stage is essentially at the beginning, when deciding what and how data can be collected while dealing with historical materials. These items can be documents, artifacts, and other kinds of information-bearing objects. They might be digitized or not-digitized, be textual or non-textual, and exist in all kinds of formats and media. For those researchers in need of historical data that cannot be obtained through web crawling or scraping, the data and information resources provided by LAMs have extraordinary value. The last two deca-

Acknowledgements

The author would like to thank Dawn Sedor for providing valuable feedback and editorial assistance.

des have witnessed a huge investment in digitizing, documenting, and making LAM resources accessible online. In order to enhance LAM data's quality and discoverability while enabling a self-sustaining ecosystem, "semantic enrichment" becomes a strategy increasingly used during recent years.

This article introduces a number of semantic enrichment methods and efforts that can be applied to LAM data at various levels. After the primer explanation of a set of key concepts, the key methods and approaches are explained through the types of data to be enhanced, mainly categorized as structured, semi-structured, and unstructured data. More specifically, these include such methods as: enhancing existing metadata's quality and discoverability with more contextualized meanings by using knowledge organization systems (KOS) vocabularies and other resources that have embraced Linked Open Data (LOD); transforming semi-structured data into structured data through entity-based semantic analysis and annotation to extend access points; digging into unstructured data and generating knowledge bases to support knowledge discovery; enabling one-to-many usages of LAM data across data silos while delivering intuitive user interfaces online; and making the heterogeneous contents from diverse providers semantically interoperable via shared ontology infrastructure. Each section ends with a discussion of representative approaches and additional resources devoted to semantic enrichment. The article concludes with the benchmarks recommended by the *W3C* (2017) in *Data on the Web best practices* which identify the ultimate goals for LAM data: comprehension, processability, discoverability, reuse possibility and effectiveness, trustiness, linkability, accessibility, and interoperability.

2. Key concepts

This review article focuses on semantic enrichment for enhancing LAM data and supporting Digital Humanities. Several key concepts need to be explained before the key approaches and methods are introduced.

Digital humanities (DH) have commanded increasing attention worldwide over the past several years. Although the definitions are being debated and the multifaceted landscape is yet to be fully understood, most agree that initiatives and activities in DH are at the intersection of the disciplines of the humanities and digital information technology. It is at this junction where digital technology will generate a paradigm shift in the near future, enabling scholars to identify major patterns in history, literature, and in the arts. DH refers to new modes of scholarship and institutional units for collaborative, transdisciplinary, and computationally engaged research, teaching, and publication (**Svensson**, 2010; **Burdick** *et al.*, 2012; **Van-Ruyskensvelde**, 2014). It is important to point out that the mere use of digital tools for the purpose of humanistic research and communication does not qualify as DH; nor is DH to be understood as the study of digital artifacts, new media, or contemporary culture in place of physical artifacts, old media, or historical culture (**Burdick** *et al.*, 2012).

One example of the growing DH movement is the *Digging into Data Challenge* program, which has funded dozens of projects aimed at research questions in the humanities and/or social sciences.
https://diggingintodata.org

Key concepts expressed in the project descriptions of *Digging into Data Challenge* Round 1-4 (2009, 2011, 2013, 2016)

Domains / Areas of Interests	Resources	Approaches
activities in humanities & social science	audio (music) recordings	annotation
ancient language	cuneiform tablets (Mesopotamia)	comparative analysis
archaeology	folklore collections	computational analysis
biodiversity	GDP per capita	computing
child language development	geographical data	corpus building
colonization of America	GitHub	cross datasets analysis
comparative and epidemiological paradigms	journals	cross-datasets searching
criminal intent	knowledge graphs	cross-linguistic annotation
debating	knowledge organization systems	data management
early modern common placing	letters	data mining
economics	linguistics databases	image processing
English speech	manuscripts	indexing
epidemiology	manuscripts (pre-modern European)	linking
film and media history	maps	machine coding
financial systems	medical images	machine learning
history	medieval charters	machine translation
human migration	multilingual classic text	metadata aggregation
human rights violations	music info	metadata analysis
information networks	news about terrorism	metadata auto-generation
information patterns and behaviors	newspapers	metadata extraction
journalism	open access publications	natural language processing (NLP)
language evolution	papyrus documents	protocols development
legal structures	passages	spatial-temporal correlation
linguistics	poetry	speech mining
literary networks	population databases	text analysis
manuscripts provenance	proceedings	visualization
music	quotations	
musicology	records in indigenous style	
parliaments	records in Spanish	
policy	signs	
population	social media	
railroad	speech datasets	
social sceince	speech recordings	
sociological theory	speeches	
standards of living	spoken language collections	
storytelling traditions and story repertoires	tweets, political	
trading and financial markets	video data	
vocabularies	writing pieces	

-Source: Compiled based on the short descriptions available at https://dev.diggingintodata.org/awards

Figure 1. Key concepts expressed in the project descriptions of *Digging into Data Challenge*.
Source: Created by the author based on project descriptions available at:
https://diggingintodata.org/awards

Since 2009, the number of participating funding organizations and nations has expanded dramatically.[1] Based on the project descriptions of the four rounds (2009, 2011, 2013, and 2016), the resources (see central column "Resources" in Figure 1) vary widely, ranging from unstructured data assets originating in ancient times to structured datasets created in the digital age. Methodologically, the projects are interdisciplinary and strive to show how best to tap into data in large-scale and diverse formats in order to search for key insights, while also ensuring access to such data by researchers through new technology-supported tools. These approaches demonstrate the essential efforts to semantically enrich the data. [Figure 1]

DATA is a concept that needs to be agreed upon when putting data into the context of digital humanities. In the digital age, it is common for people to think of data only in terms of digitally available formats. The connection between digital data and data analytics is correct, but we need to fully understand that the terms "data" and "digital data" are not equivalent. Types of data are also not limited to quantitative data.

The *Reference model for an Open Archival Information System* (*OAIS*) defines data as a

> "reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing"

while offering examples of data as: a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. This definition of "data" was given within the context of "information," which is

> "Any type of knowledge that can be exchanged. In an exchange, it is represented by data" (*Consultative Committee for Space Data Systems*, 2012, 1-10 and 1-12).

In the book *Information: A very short introduction*, Luciano Floridi defines data at its most basic level as the absence of uniformity, whether in the real world or in some symbolic system. Only once such data have some recognizable structure and are given some meaning can they be considered information (**Floridi**, 2010, pp. 22-25).

After a comprehensive review of the definitions and terminology for "data" in her book titled *Big data, little data, no data: Scholarship in the networked world*, Christine Borgman summarized that

> "data are representations of observations, objects, or other entities used as evidence of phenomena for the purpose of research or scholarship" (**Borgman**, 2015, p. 28).

LAM DATA is a broad term used in this article to refer to all data provided by LAMs and other information institutions. They provide tremendous opportunities for humanities researchers to unearth nuggets of gold from the available data. LAM data can be categorized in three main groups based on their type:

- *Structured data* in LAMs include bibliographies, indexing and abstracting databases, citation indexes, catalogs of all kinds, special collections portals, metadata repositories, curated research datasets, and name authorities. Structured data are typically held in databases in which all key/value pairs have identifiers and clear relations and follow an explicit data model (**Schöch**, 2013).
- *Semi-structured data* in LAMs comprise the unstructured sections within metadata descriptions (e.g., notes in bibliographic records, the rich content descriptions contained in archival finding aids, table-of-contents and abstracts of reports in digital repositories), archival documentation not carried in *Encoded Archival Description* (EAD) or other digital finding aids, intellectual works encoded following the *Text Encoding Initiative* (TEI) guidelines (main text body excluding the Header), value-added or tagged resources that exist in all kinds of formats, and the unstructured portions of otherwise structured datasets.
- *Unstructured data* are

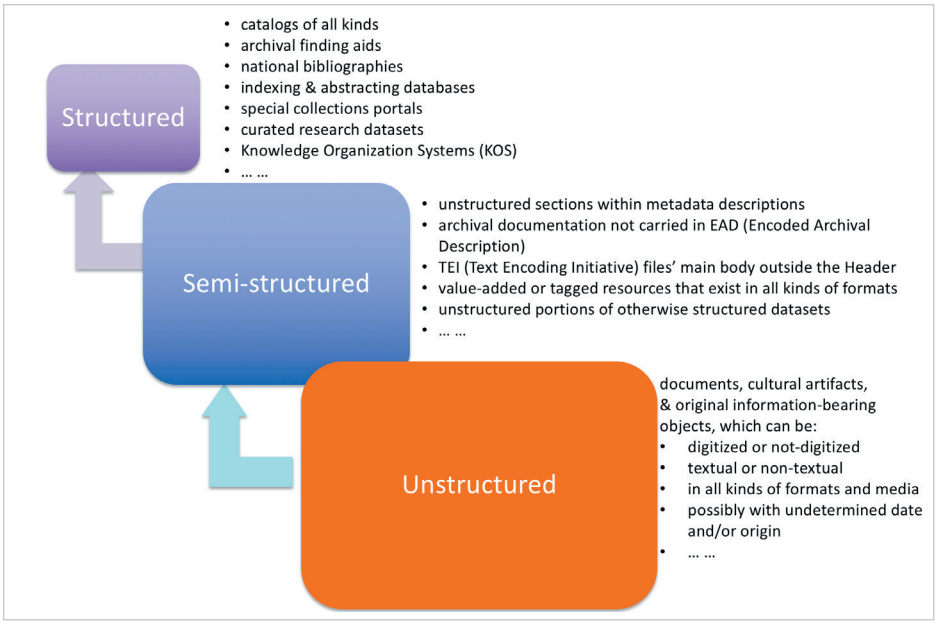


Figure 2. LAM data examples.

characterized as "everything else." In LAMs, they can be found in documents and other information-bearing objects (textual or non-textual, digitized or non-digitized), in all kinds of formats. In data resources that are served through LAMs, unstructured data are usually available in the largest quantity in comparison with structured data, have the most diversity in type, nature, and quality, and are the most challenging to process. [Figure 2]

The primary LAM data assets might be held in special collections, archives, oral history files, annual reports, provenance indexes, and inventories, to name just a few. During the last two decades, many digital collections were born in LAMs and have exhibited strong outcomes. These projects have digitized and delivered integrated resources (metadata, representative images, original documents, and media) on the web. The creation of these digital products involved very complicated digitization and documentation processes requiring tremendous investments from government and other institutions. For example, from 2002 to 2011, in the United States, the federal agency *Institute of Museum and Library Services' IMLS Grants to States* program supplied $980 million to support increased access to digital information, including $67 million toward the digitization of local history and special collections (*IMLS*, 2018, p. 9). Having invested so much, it is critical for LAMs to extend the values of these digital collections beyond being just retrospective resource warehouses in order to have them be better shared, linked, enriched, and reused.

Thus, the next challenge is to move from digitizing to datafying. To "datafy" the unstructured data means to turn the heritage materials into not only machine-readable but also machine-processable resources, and reconstruct materials through digitization pipelines. The demand for datafication might explain why, for digital humanities, the Smart Data approach emphasizes the processes to transform unstructured data to structured and semi-structured data (**Schöch**, 2013; **Mayer-Schönberger**; **Cukier** 2013; **Kaplan**, 2015).

SMART DATA is a concept embraced by humanities research, and underlines the organizing and integrating processes from unstructured data to structured and semi-structured data, making the big data smarter (**Kobielus**, 2016; **Schöch**, 2013). The concept should be understood in the context of Big Data. Among the many "V"s of Big Data (volume, velocity, variety, variability, veracity, and value), the "V"alue of data relies on the ability to achieve big insights from such data on any scale, great or small (**Kobielus**, 2016).

> "[I]n its raw form, data is just like crude oil; it needs to be refined and processed in order to generate real value. Data has to be cleaned, transformed, and analyzed to unlock its hidden potential." (*TiECON East*, 2014).

Only after it has been tamed through the organization and integration processes is such data turned into Smart Data that reflects the research priorities of a particular discipline or field. As Smart Data inquiries, these tamed results can then be used to provide comprehensive analyses and generate new products and services. (**Gardner**, 2012; **Mukerjee**, 2014; **Schöch**, 2013). [Figure 3]

Schöch concluded in his monumental article *Big? Smart? Clean? Messy? Data in the humanities* that we need Smart Big Data because it both adequately represents a sufficient number of relevant features of humanistic objects of inquiry to enable the necessary level of precision and nuance required by humanities scholars, and it provides us with a sufficient amount of data to enable quantitative methods of inquiry, helping us surpass the limitations inherent in methods based on close reading strategies (**Schöch**, 2013). Researchers in the humanities have incorporated the data-driven environment where advanced digital technologies have created the possibility of novel and hybrid methodologies. In the processes that transform unstructured data to structured and semi-structured data, the Smart Data strategy drives data service providers to aim at supporting DH by:

- creating machine-understandable, -processable, and -actionable (instead of merely machine-readable) data;
- providing accurate data in the processes of interlinking, citing, transferring, rights-permission management, use and reuse;
- enabling both one-to-many usages and high efficiency processing of data (**Zeng**, 2017).

The SEMANTIC ENRICHMENT strategy represents one of the major steps to align with the aim of Smart Data. Semantic enrichment is directly applied

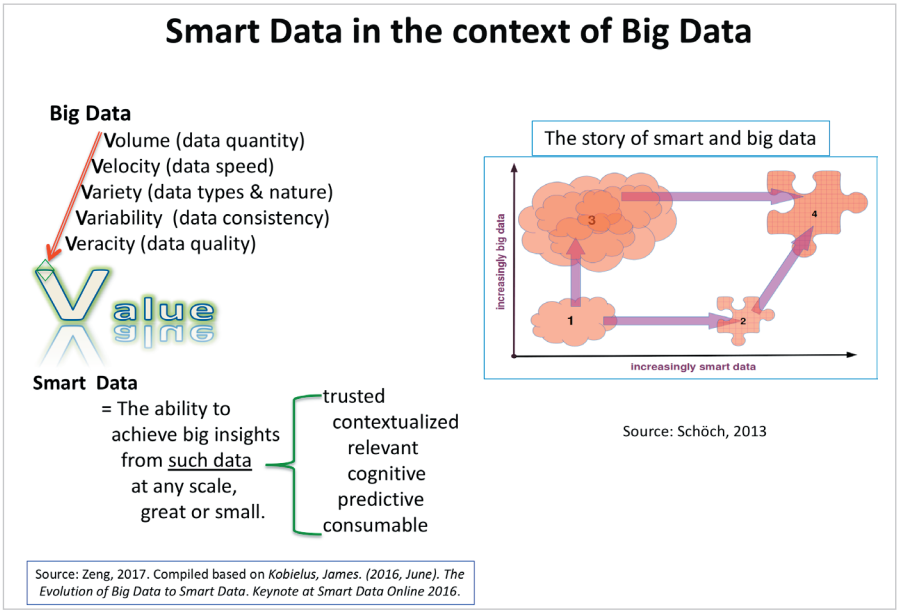


Figure 3. Smart Data in the context of Big Data.

to the enhancement of LAM data (structured, semi-structured, and unstructured) by using semantic technologies. In a broad sense, semantic enrichment in the context of data may aim at different targets. **Damjanovic** *et al.* (2011) surveyed various semantic enhancement approaches and techniques and presented four categories.

- *Semantic search and browsing*, which includes distinct research directions such as (i) augmenting traditional keyword search with semantic techniques, (ii) basic concept location (e.g. multi-facet search, semantic auto-completion, search behavior research), (iii) complex constraint queries for creating query patterns as intuitively as possible, (iv) problem solving, and (v) connecting path discovery.
- *Semantic mediation: merging and mapping*. Merging unifies two or more ontologies with overlapping parts into a single ontology that includes all information from the sources. Mapping builds the mapping statements that define relationships between concepts of ontologies and rules that specify transformations between two ontologies.
- *Semantic annotation*, which formally identifies concepts and relations between concepts in documents, and is intended for use by machines.
- *Semantic analytics and knowledge discovery*. Semantic analytics is a process of analyzing, searching, and presenting information by using explicit semantic relationships between known entities. Both federated and centralized approaches to processing queries on Linked Open Data have been used. (**Damjanovic** *et al.*, 2011)

The article also anticipates the revolutions of Web-based Content Management Systems (WCMS), which replaces in-house-developed CMS for intranet sites and integrates firmly within the Web and document-oriented environments.

Also addressing the differences between knowledge organization in the bibliographic domain and requirements for resource discovery in a web environment, **Prasad**, **Giunchiglia** and **Devika** (2017) presented the *DERA* model (Domain, Entity, Relations, Attributes) featuring the transition from document-centric to entity-centric knowledge modeling. The authors believe that any domain following *DERA* can be seen as formalized by the structure *D<E,R,A>*, where:

- Domain (D) is a particular area of knowledge or field of interest being studied.
- Entity (E) consists of a set of facets where each facet represents a group of terms denoting the entity classes of the real-world entities (instances) having perceptual correlates or only conceptual existence.
- Relation (R) consists of a set of facets where each facet represents a group of terms denoting the relations between entities. Each relation term establishes a semantic relation between two entities.
- Attribute (A) consists of a set of facets where each facet represents a group of terms denoting the qualitative and/or quantitative properties of entities (**Prasad**; **Giunchiglia**; **Devika**, 2017).

It is important to refer to the *International Federation of Library Associations and Institutions* (*IFLA*)'s new *Library Re-*

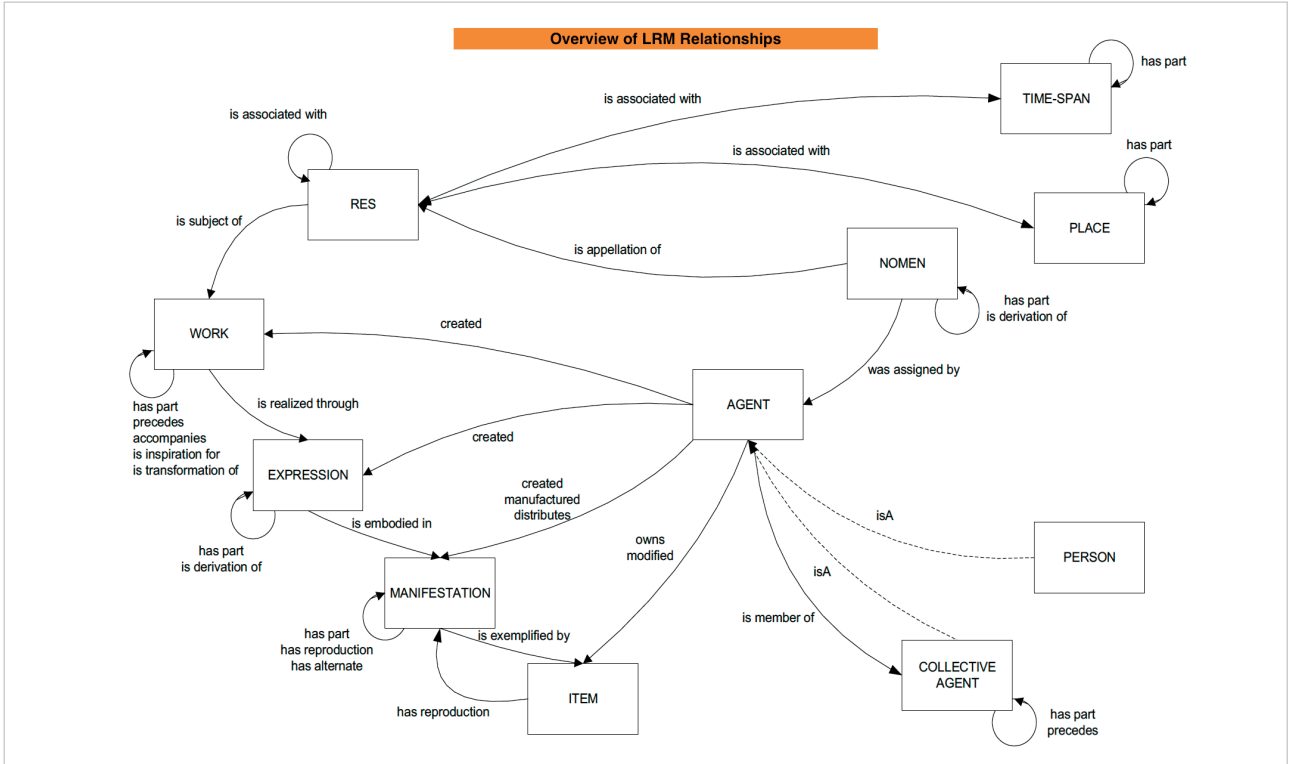


Figure 4. Overview of *IFLA Library Reference Model* (*LRM*) relationships.
Source: **Žumer** and **Riva**, 2017.

ference Model (*IFLA LRM*) in this obvious transition from document-centric to entity-centric modeling trend. *IFLA LRM* consolidated the three models of the *FRBR* Family (*FRBR*,[2] *FRAD*,[3] and *FRSAD*[4]) and *LRM* was formally adopted by the *IFLA Professional Committee* in August 2017 (**Riva**; **LeBoeuf**; **Žumer**, 2017). *LRM* has presented a complete model of the bibliographic universe using entity-relationship modeling. In contrast with the *FRBR* Family, where all entities are at the same level, a hierarchical structure of entities is introduced in *LRM* by declaring entities within a structure of superclasses and subclasses. That one entity is a subclass of another entity can be expressed using the ***isA*** relationship. For example, *Agent* (an entity capable of deliberate actions, of being granted rights, and of being held accountable for its actions) is declared as a superclass which brings a hierarchical structure to handle the former *FRBR Group 2* of entities. This powerful mechanism enables considerable simplification of the model, because attributes and relationships can be declared on the higher level and do not have to be repeated on lower levels. *Res* (any entity in the universe of discourse) is the superclass of all entities in the model. *Nomen* (an association between an entity and a designation that refers to it) is an entity itself, being the appellation used to refer to an instance of *Res*. In order to model more precisely the temporal and spatial aspects, *LRM* introduces two additional entities, *Place* and *Time-span*. All of these entities presented by *LRM*, as well as the more semantically meaningful attributes and relationships, make *LRM* foundational for the development of cataloguing rules and bibliographic formats (**Žumer**, 2018; **Žumer**; **Riva**, 2017). [Figure 4]

IFLA LRM defines five user tasks (Table 1) and lists the goals users want to reach when performing the tasks. The term resource is used in its broadest sense, meaning an instance of any entity defined in the model. The tasks of *find*, *identify*, *select*, and *obtain* are the same as defined in *FRBR*, with slightly modified and broader definitions; *explore* was first introduced by *FRSAD* (**Žumer**, 2018).

Table 1. A summary of user tasks defined by *LRM*

Find	To bring together information about one or more resources of interest by searching on any relevant criteria
Identify	To clearly understand the nature of the resources found and to distinguish between similar resources
Select	To determine the suitability of the resources found, and to be enabled to either accept or reject specific resources
Obtain	To access the content of the resource
Explore	To discover resources using the relationships between them and thus place the resources in a context

From the perspective of user tasks, it is clear that most of the metadata semantic enrichment efforts extend the initial functions of bibliographic control and enable the *explore* task to be accomplished meaningfully and effectively.

Semantic enrichment process pertaining to LAM data reflects the transformation from document-centric to entity-centric knowledge modeling. The process is distinguished by three main stages in the *Europeana Semantic Enrichment Framework* documentation:

- Analysis: the pre-enrichment phase focuses on the analysis of the metadata fields in the original resource descriptions, the selection of potential resources to be linked to, and derives rules to match and link the original fields to the contextual resource.
- Linking: the process of automatically matching the values of the metadata fields to values of the contextual resources and adding contextual links (whose values are most often based on equivalent relationships) to the dataset.
- Augmentation: the process of selecting the values from the contextual resource to be added to the original object description. This might not only include (multilingual) synonyms of terms to be enriched but also further information, for example broader or narrower concepts (**Isaac** *et al.*, 2015, pp. 8-9; **Manguinhas**, 2016).

> “Data is a concept that needs to be agreed upon when putting data into the context of digital humanities. The connection between digital data and data analytics is correct, but the terms “data” and “digital data” are not equivalent”

The “contextual resources” referred by this documentation and the real cases to be introduced in the next chapter of this article (Section 3.1.1.) signify the selected vocabularies and datasets. For all types of vocabularies and schemes for organizing information and promoting knowledge management, “knowledge organization systems (KOS)” is a broad term that can be used. Examples of KOS include pick lists, authority files, gazetteers, synonym rings, taxonomies and classification schemes, lists of subject headings, thesauri, and ontologies. An instance might be referred to as a “controlled vocabulary,” a “value vocabulary,” or, in a broader sense, a “taxonomy,” by different communities. In this article, when an instance of knowledge organization systems is mentioned, the term KOS vocabulary will be used. For those KOS vocabularies that have been published as Linked Open Data (LOD), they will be referred to as LOD KOS (**Zeng**; **Mayr**, 2018). In the semantic enrichment projects, LOD KOS and other LOD datasets are essential.

Core agents in the semantic enrichment processes are explained according to the concepts identified in the report of the *Europeana Task Force on Enrichment and Evaluation* (**Isaac** *et al.*, 2015). Even though the project focused on structured data, the concepts can also be applied to a wide range of semantic enrichment of LAM data, structured or not.

- At the beginning of each enrichment process, there is the source data that will be enriched. This data comes from different data providers.
- The agent in charge of selecting the different datasets and processing them for enrichment is the enricher - the one who handles the process of enrichment.
- The user of the services made possible or enhanced through enrichment is the end user.
- Sometimes, enrichments can be crowdsourced. In these cases, the volunteers using the crowdsourcing tool and annotating data are the annotators.

3. Key approaches / Methods for LAM data's semantic enrichment

After getting the key concepts explained above, this chapter presents the key approaches and methods for LAM data's semantic enrichment, using cases and experiments reported in the 2010s. These will be explained through the types of data to be enhanced, mainly categorized as structured, semi-structured, and unstructured data.

3.1. Semantic enrichment of structured data

Enriching structured data (often referring to metadata) has become a common initiative in LAM data enhancement efforts, in order to overcome challenges relating to data quality and discoverability in the digital age, while providing more context and multilingual information for cultural heritage (CH) objects. The term "enrichment" may refer to the process, e.g., the application of an enrichment tool, or to its results, such as the new metadata created at the end of the process (**Isaac** *et al.*, 2015). In literature, various terms may reference such methods as reconciliation, mapping, alignment, matching, massaging, merging, interlinking, etc. The overall result is clearly the enrichment of existing metadata, with more contextualized meanings.

Methods and cases discussed in this section have a starting point: the structured LAM data, especially the components where data values are in a controlled /normalized form and are expected to be the access points. As demonstrated by these cases, such data values can reside in original metadata descriptions for CH objects in a LAM data digital platform (e.g., metadata in *Europeana* and *Swissbib*), or exist in agent-centered information clusters (e.g., web pages delivered by the *Museum of Modern Art* (*MoMA*) and the website of the *Museums and Collections with Maya Inscriptions*). The initial "source" and "target" substances involved in the alignment process can be any of these types: metadata descriptions (e.g., *Europeana* metadata), KOS vocabularies and other contextual resources [e.g., *GeoNames*, *VIAF*, *Faceted Application of Subject Terminology* (*FAST*), *Wikidata*, *DBpedia*, etc.], or information resources (e.g., *Wikipedia* entries, biographies, geo-maps) where the focused subjects are the entities in metadata descriptions or KOS vocabularies. Their positions in the alignment (as a "source" or a "target") and the directions of linking can be switched according to a project's design.

3.1.1. Example: structured data in original metadata descriptions for CH objects

Case: *Europeana*

Considered first as an experiment, metadata enrichment has become part of the strategy of *Europeana* and its data providers. The semantic enrichment intends to enrich data providers' metadata by automatically linking text strings found in the metadata to contextual resources from selected LOD datasets or KOS vocabularies. (**Stiller** *et al.*, 2014; **Isaac** *et al.*, 2015) The method comprises augmenting the source metadata with additional terms, seen in two steps:

- Matching the metadata of *Europeana* objects to external semantic data results in links between these objects and resources from external datasets.
- The created links point to additional data such as translated labels or broader labels. A record might be supplemented with all the translated labels of the *DBpedia* concept as well as linking to a broader concept in *DBpedia* and all its translated labels. (**Isaac** *et al.*, 2015)

An example of these processes and results are provided on the *Europeana* semantic enrichment website[5] and demonstrated in Figure 5 in the subsection below.

This automatic linking method is effective when applied to those metadata values that are in a controlled form, including place, agent, concept, and time period. For instance, it may enrich the names of places with values from *GeoNames*, while person names and concepts are enriched with values from *DBpedia* and other vocabularies. The pattern can be simply interpreted as:

"*Europeana* enriches names of … with values from [xxx]", where … and [xxx] can be:

- PLACE [*GeoNames*, *Gemeinsame Normdatei* (*GND*)]
- AGENT [*Virtual International Authority File* (*VIAF*), *The Getty Union List of Artist Names* (*ULAN*), *MIMO Instrument makers*, *GND*, *DBpedia*, etc.]
- CONCEPT [*Art & Architecture Thesaurus* (*AAT*), *Unesco Thesaurus*, *WWI Concepts* from *Library of Congress Subject Headings* (*LCSH*), *Universal Decimal Classification* (*UDC*), *MIMO Concepts*, *IconClass*, *GND*, *Europeana Sounds Genres*, *DBpedia*, etc.]
- TIME PERIOD [*Semium Time*]

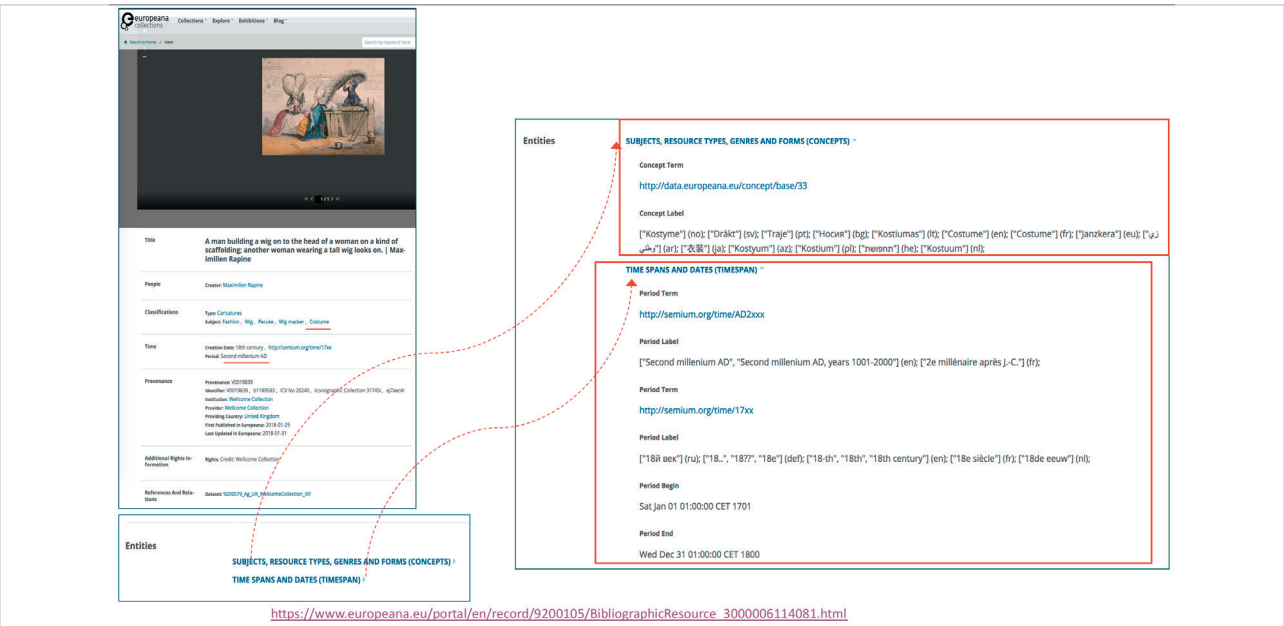


Figure 5. Example of a *Europeana* record semantically enriched with additional data such as translated labels.
Source: Image captured from *Europeana* website at
https://www.Europeana.eu/portal/en/record/9200105/BibliographicResource_3000006114081.html

Those [xxx] represent the external datasets or KOS vocabularies currently chosen by *Europeana* for the automatic alignment process. They are internationally established initiatives or more specific projects whose vocabularies are published as LOD (refer to list of *Europeana Dereferenceable vocabularies*).[6] *Europeana* has developed a tool that 'dereferences' the URIs, i.e., that fetches all the multilingual and semantic data that are published as LOD for vocabulary concepts and other contextual resources on third-party services. *Europeana* encourages participants to use them while also accommodating the participant's own LOD KOS vocabulary/vocabularies.

An enrichment may manifest links that were already implicitly present in the data, as in the case of metadata 'massaging' in *Europeana*. It can be done through advanced mapping or by using tools such as *OpenRefine* where the (string) label of a concept used in an object's metadata is replaced by the identifier of this concept used in its original KOS vocabulary (**Isaac** *et al.*, 2015, p. 11.) Figure 5 shows additional data, such as multilingual labels of the concept from external KOS vocabularies, that result from the enriched metadata section of Entities.
http://openrefine.org

Great progress has led to millions of semantically enriched metadata. Based on experiments conducted around 2015, *Europeana* performed quantitative and qualitative evaluations of seven enrichment services on the same subset of a *Europeana* dataset containing 17300 records. The quantitative overview of the results of the semantic enrichment have been updated by the team (**Manguinhas**, 2016). Millions more have already been added for concepts, places, agents, and time spans since 2017.[7]

Contextualization implies creating "typed relationships" between resources of different types. This process of contextualization involves matching between two objects, two places, or two concepts, e.g., considering whether the match of two concepts are *exactMatch* or *closeMatch*, whether a concept is *broader* or *narrower* than the one aligned to, or whether the two identifiers from two sources actually represent the same place. The *Europeana Data Model* (*EDM*) is the core for the defined properties which express various types of relationships in the alignment results. The Figure 6 illustrates some of the types of links specified. [Figure 6]

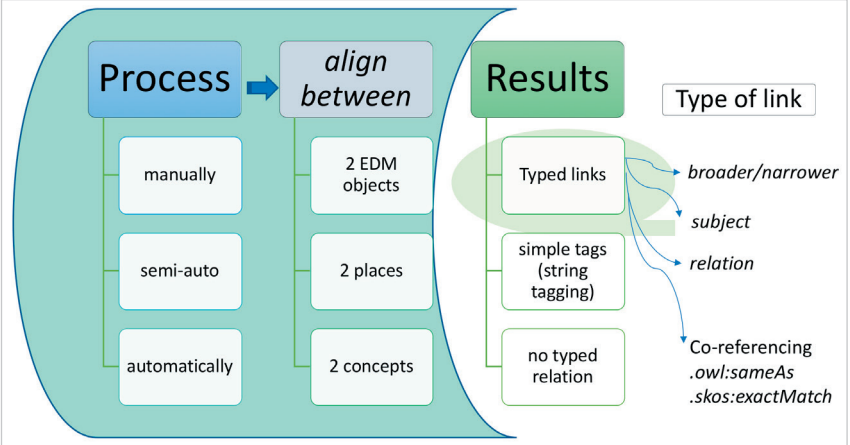


Figure 6. Contextualization - creating typed relationships between resources of different types.
Source: Image based on **Isaac** *et al.*, 2015.

Case: *Swissbib*

The enrichment of structured data may be centralized on a specific type of entity, for example, authors. *Swissbib*, a provider for bibliographic data in Switzerland, reported a process for extracting and shaping the data into a more suitable form (see Figure 8). In one example, data available in MARC21 XML were extracted from the *Swissbib* system and transformed into an RDF/XML representation. In another exercise, author information was extracted from approximately 21 million monolithic records and interlinked with authority files from the *VIAF* and *DBpedia*. [Figure 7]
http://linked.swissbib.ch

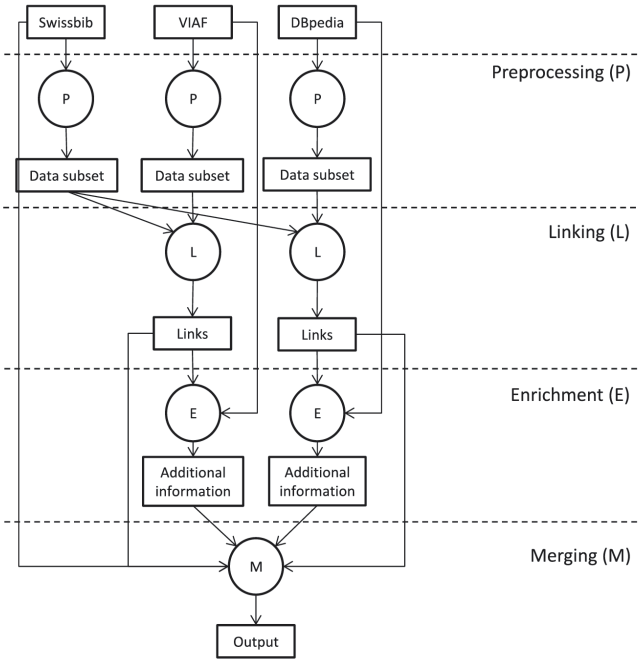


Figure 7. Data flow diagram of the interlinking procedure in the *Swissbib* project
Source: **Bensmann**; **Zapilko**; **Mayr**, 2017, p. 8.

In the enrichment process, the researchers take the links <swissbib> <owl:sameAs> <viaf>, sort them by their object and align them with the respective external corpus. In this way, selected statements about the referenced author are extracted from the external corpus and rewritten to make them statements of the *Swissbib* author resource. Particular statements may refer to further resources instead of literals, e.g., locations. In order to be able to display these resources on the user interface, these resources are summarized into a single literal that represents the resource in a suitable manner, e.g. using labels or descriptions. The resulting literal is additional to the original property, added to the person description using a new extended property (*dbp:birthPlace->swissbib:dbpBirthPlaceAsLiteral*). Finally, all persons, together with the links and extracted data, are deposited at an agreed location for indexing. The approach established 30,773 links to *DBpedia* and 20,714 links to *VIAF*, and both link sets show high precision values and were generated in reasonable expenditures of time, according to the authors (**Bensmann**; **Zapilko**; **Mayr**, 2017).

3.1.2. Example: structured data in agent-centered information clusters on the Web

Case: *Museum of Modern Art* (*MoMA*)

A unique named-entity-centered case is the information cluster on website of the *Museum of Modern Art* (*MoMA*) in New York City, United States. The *Museum*'s website features 72,706 artworks from 20,956 artists (*MoMA*, 2017). In addition to web pages, two types of datasets are openly available:[8]

(1) The artworks dataset (artworks.csv) contains more than 130,000 records, including basic metadata for each work, such as title, artist, date, medium, dimensions, and date acquired by the *Museum*.

(2) The artists dataset (artists.csv) contains more than 15,000 records, representing all artists who have work featured in *MoMA*'s collection and have been cataloged in the database. The structured data for each artist includes name, nationality, gender, birth year, and death year. By mapping the artists dataset to the *Union List of Artist Names* (*ULAN*) data through *Getty Vocabularies*: LOD,[9] the majority of the artists in the *MoMA* dataset obtained http URIs from the *ULAN*. [Figure 8]

From the front-end of the webpage which shares information about the artist as well as related exhibitions (Figure 8, center), it is clear that the artist's information is greatly enriched by the high-quality *ULAN* name authority data, which include the information about an artist, such as: name, bio, nationality, role, type, and multilingual appellations (Figure 8 right). If one chooses to "View the full Getty record"[10], a user can find not just identity information, but also rich data, such as: the artist's associative relationships with related people or corporate bodies, sources of relevant information, aligned entries with other name authorities, and many publications and databases where the artist is the subject.

> With the rapid development of the digital humanities field, demands for smarter and bigger historical and cultural heritage data, which usually cannot be obtained through web crawling or scraping, have generated deep interest in LAM data, the treasure of all society. The semantic enrichment strategy represents one major step and directly enhances LAM data by using semantic technologies

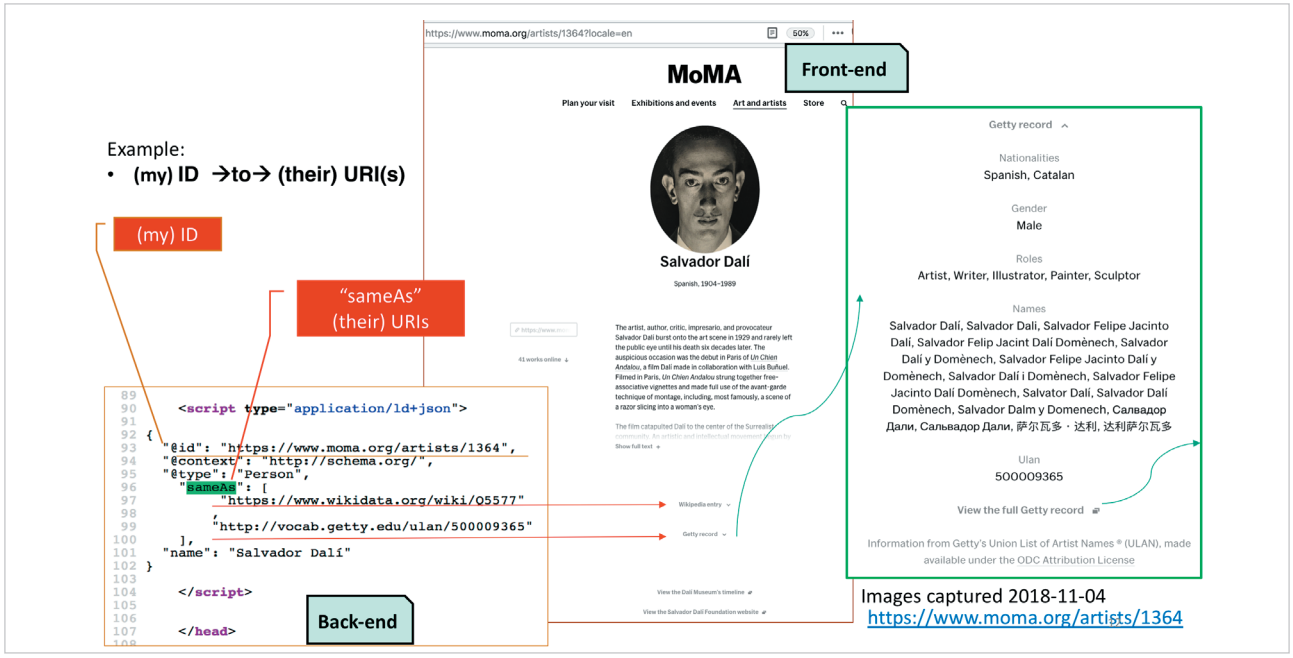


Figure 8. Front-end and back-end of the "massaged" results, using a *MoMA* artist webpage as an example.
Source: Generated by the author based on *MoMA* webpage and its source code view.
https://www.moma.org/artists/1364

The aligned contents are presented not just in the front-end webpages (explained above). At the back-end (Figure 8 left), the artist's URIs from *Wikidata* and *ULAN* are embedded in the <head> section of the HTML page, coded with schema.org property ***sameAs***. The impact of such an enrichment also exposes resources about the focused entity of interest (e.g., the artist) to the web. For example, searching the artist's *ULAN* identifiers through search engines (e.g., *Google*) would bring highly relevant links on the first two results pages. In this way, the ranks of significant cultural heritage institutions such as museums and libraries are actually greatly increased, placing them ahead of hundreds of other results pushed out by search engines.

It has become a common method to use tools such as *OpenRefine* to consolidate unstructured data, where the (string) label of a concept used in an object's metadata (e.g., *MoMA*'s artist records available in a CSV file) is aligned with the identifier of this concept in its source KOS vocabulary (e.g., *ULAN*). In the *MoMA* case, the process is described in the guidelines provided by the *Getty Vocabularies LOD service*.[11]

Case: A checklist of ***Museums and Collections with Maya Inscriptions***

A similar agent-centered information cluster resource is the *Museums and Collections with Maya Inscriptions* website, developed by the *Interdisciplinary Dictionary of Classic Mayan* (*Textdatenbank und Wörterbuch des Klassischen Maya*) research center at the *University of Bonn*.[12] One of the corpus databases has been constructed for objects that are now housed in museums and collections. The website provides a resource listing all museums and collections with Mayan inscriptions worldwide. Each page provides the museum or collection's name, location, contact information, and links to the museum's website, catalogs and databases. When possible, identifiers from *GeoNames*, the *Getty Thesaurus of Geographic Names* (*TGN*), and the *Union List of Artist Names* (*ULAN*) are included. A map (*Google Maps*) showing the exact location of the relevant museum or collection is also given for

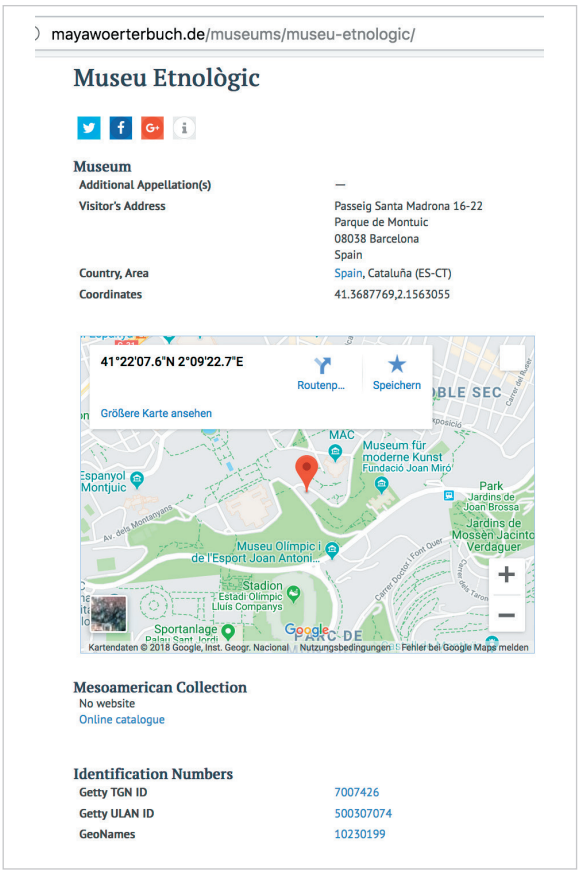


Figure 9. A webpage about a museum including the "massaged" results including the identifiers of *Getty TGN*, *ILAN*, and *GeoNames*.
Source: Entry of *Museu Etnològic*
http://mayawoerterbuch.de/museums/museu-etnologic

each record (**Wagner** *et al.*, 2014). For example, the page for *Museu Etnològic* in Barcelona, Spain[13] provides a direct link to *ULAN* ID 500307074[14] for the museum's name authority record, the *Getty TGN* ID 7007426[15] for Barcelona, and *GeoNames* 10230199[16] for *Museu Etnològic*, plus a geographic map showing the exact location of the relevant museum or collection. [Figure 9]
http://vocab.getty.edu/page/ulan/500307074
http://vocab.getty.edu/page/tgn/7007426
http://www.geonames.org/10230199/museu-etnologic.html

> Enriching structured data has become a common initiative in LAM data enhancement efforts. These structured data have been developed and sustainably maintained; they exist and are ready to be enriched

3.1.3. Example: structured data in subject authority entries

Case: *Faceted Application of Subject Terminology* (*FAST*)

Conversely, this kind of contextualization or "massage" can also be conducted to enrich an existing KOS vocabulary by aligning to external contextual resources. What is enriched is not a metadata record for an object, but rather the subject authority entry. The important concept to be revisited here is the typed relationships that are applied in the massaged process. *Faceted Application of Subject Terminology* (*FAST*) has used *schema:sameAs*, *owl:sameAs*, and *foaf:focus* to allow *FAST*'s controlled terms (representing instances of *skos:Concept*) to be connected to URIs that identify real-world entities specified at *VIAF*, *GeoNames*, and *DBpedia*.
https://www.oclc.org/research/themes/data-science/fast.html

In the following example, the highlighted codes indicate the relevant coding of properties that connects the artist entry in *Wikipedia* through *foaf:focus* property, while the next set of codes shows how the data from *VIAF* is connected through *schema:sameAs*.

- through *foaf:focus*, the *Wikipedia* URI allows a *FAST* concept to connect with information about the concept, which is usually excluded in authority records;
- through schema:sameAs, the identifier of *VIAF* lets *FAST* take advantage of all of the various string values included in *VIAF* (containing dozens of multilingual name authorities) without having to manually include the values in the RDF triples for the specific entry in *FAST*.

With these accurate typed relationships, machines can understand (and reason) that a *FAST*'s controlled term is related to a real-world entity, and allows humans to gather more information about the entity that is being described (**O'Neill**; **Mixter**, 2013). [Figure 10]

FAST Authority File example

Entry 39278 for artist Gaudí, Antoni, 1852-1926 is enriched through the alignment with external identifiers using typed-relationships, foaf:focus and owl:sameAs

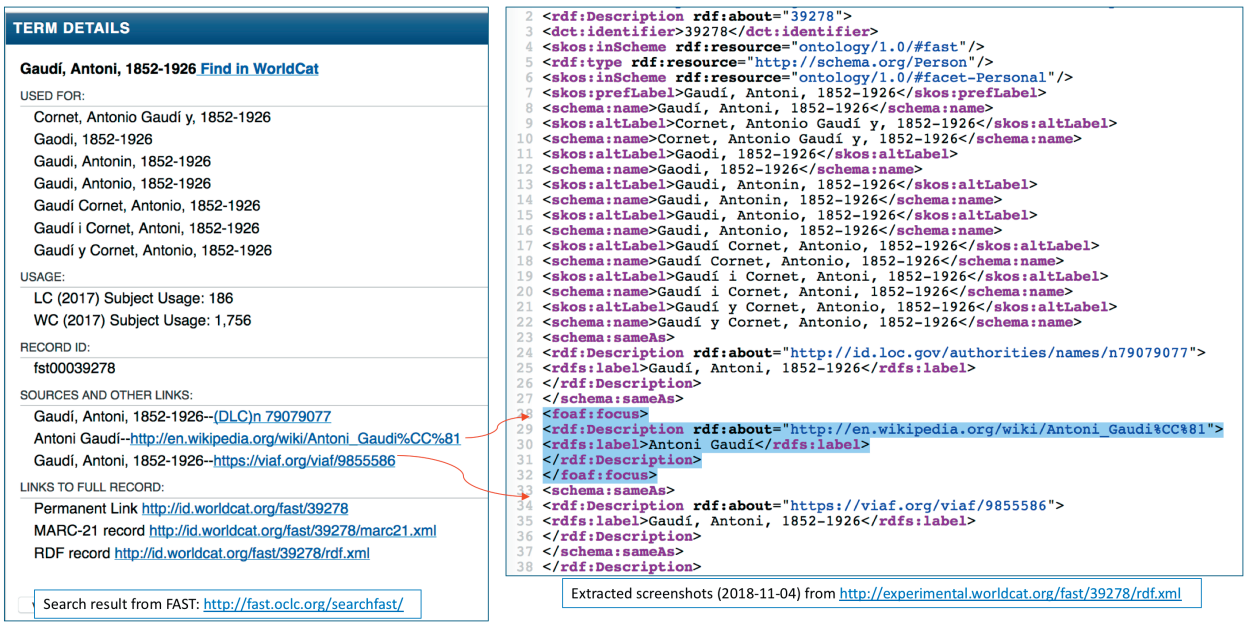


Figure 10. Examples from *Faceted Application of Subject Terminology* (*FAST*) showing the machine understandable coding of linkages to external KOS vocabularies and contextual resources.
Source: screenshots from *FAST*, 2018-11-01.

3.1.4. Discussion

Enriching structured data has become a common initiative in LAM data enhancement efforts, and more reports can be found in publications and on the web. The successful cases presented so far have proved that, when semantic enrichment is applied to the structured data that are in a controlled/normalized form, including entities for place, agent, concept, and time period, the results and impacts are significant. These structured data have been developed and sustainably maintained; they exist and are ready to be enriched. This is a major difference of such data from the others to be discussed in the following sections (on semi-structured and unstructured data), when more complicated semantic enrichment workflows (including model developing, batch processing, validating, disseminating, etc.) might need significant additional investments and resources.

The cases introduced in this section also revealed the usage of external contextual resources for enriching those controlled values in structured data. What are such resources, then? An *International Linked Data Survey for Implementers* conducted by the *OCLC Research* in 2018 reported top ranked Linked Data sources:
1) *id.loc.gov*
2) *VIAF*
3) *DBpedia*
4) *GeoNames*
5) *Wikidata*.

In comparison with the previous 2015 survey, the biggest change was the rise in *Wikidata* as a linked data source. *Wikidata* was elevated to the 5th-ranked data source utilized by linked data projects/services in the 2018 survey, compared to 15th in the 2015 survey. Eighty-one institutions responded to the 2018 survey, describing 104 linked data projects (**Smith-Yoshimura**, 2018).

In addition to these commonly used resources, other well-established LOD KOS vocabularies are similarly invaluable for specialized areas. The cases introduced in this section demonstrated the use of some of them. On a larger scale, a substantial activity that should be noticed is *Mix'n'Match* which has brought the largest mash-up effort forward in order to fully use established, reliable vocabularies and datasets.
https://tools.wmflabs.org/mix-n-match/#/

As a tool, *Mix'n'Match* lists entries of hundreds of external databases in a variety of categories and allows volunteers to manually match them against *Wikidata* items. An exceptional feature of this resource is the number and variety external datasets: for example, dozens in the Heritage category and over 500 in the Biography category are all sourced from different countries. The *Authority control for people* has reached over 480 catalogs as of the end of 2018, including, for example, *Comité Olímpico Argentino* (*Argentinian Olympians*), *RANM* (identifiers of members of the *Spanish Royal Academy of Medicine*), *Royal Swedish Academy of Letters*, Who's *Who in France*, *Database of Classical Scholars*, etc. Taking the general "Authority control" datasets (over 100) as another example, it includes well-known vocabularies such as *GeoNames*, *FAST*, *Unesco Thesaurus*, and *MeSH* (*Medical Subject Headings*) sub-lists, as well as other specialized vocabularies such as *DoS* (*Dictionary of Sydney*), *Inran Italian Food Nutrient profiles*, *ISO 15924* numeric code, *Gran Enciclopèdia Catalana*, *Europeana Fashion Thesaurus*, *MIMO Music Instruments*, *Great Russian Encyclopedia*, etc. More than half of these vocabularies have over 70% of entries manually matched by contributors. These resources can be investigated when a semantic enrichment project task force starts defining the target datasets, as a part in the process of enrichment (refer to the 2nd step in the next paragraph).

The report of the *Europeana Task Force on Enrichment and Evaluation* (**Isaac** *et al.*, 2015) contains the most comprehensive and relevant guidelines for the whole workflow and project design. It recommends ten steps for developing and maintaining a successful enrichment strategy:

1. Define the enrichment goals (annotation guidelines) that will guide your enrichment strategy.
2. Choose the right components for your enrichment workflow: enrichment solution and target datasets.
3. Define the enrichment workflow.
4. Make sure your enricher has sufficient knowledge.
5. Test your enrichment workflow.
6. Assess the quality of your enrichment and have an evaluation strategy.
7. Choose the right evaluation method.
8. Apply user-initiated enrichment workflows.
9. Document your enrichment process and learnings.
10. Monitor your enrichment process and re-assess.

3.2. Semantic enrichment of semi-structured data, expanding access points

In the previous section on semantic enrichment of structured data, methods discussed are normally applied to components in metadata records where data values are available in a controlled form and are expected to be the primary access points. What about those uncontrolled, not-normalized, and free-text parts within metadata records, or the

unstructured segments of otherwise structured datasets? For example, although agent names in a music bibliographic record are available access points, especially those fields in a controlled form, the responsibility or role of the agents might be hidden in non-controlled components within the same record. Uncovering potentially valuable, yet hidden information and access points in the semi-structured data leads to another major category of semantic enrichment actions.

This section is based on research experiments and pilot studies addressing the semantic enrichment of semi-structured data, parsing data that are in non-controlled/not-normalized form, turning them into access points, and providing useful contextual information. Semi-structured data waiting to be parsed and semantically enriched include, for example, the text in certain MARC bibliographic records' fields, the summary section in a photograph collection metadata description, and the detailed descriptive information in archival finding aids like EAD (*Encoded Archival Description*) records, just to name a few. They are different from unstructured data because these semi-structured data contents are the results of information processing or documentation workflow and are recorded in metadata records. They assemble invaluable resources not represented in other structured data fields, and usually contain important contextual information. Collectively, the examples included in this section illustrate methodologies and findings resulting from the extensive exploration and analysis conducted by the investigators.

3.2.1. Example: semi-structured data within MARC bibliographic records

In **Weitz** *et al.*'s article *Mining MARC's hidden treasures: Initial investigations into how notes of the past might shape our future* (2016), the researchers at *OCLC* (*Online Computer Library Center*) reported their study on finding, interpreting, and manipulating the rich trove of data already present in MARC bibliographic records. The following sub-sections are all based on this article.

As cataloging moves from AACR2 to Resource Description and Access (RDA), and MARC 21 gains the explicit level of detail that will allow cataloging to move into a post-MARC environment, the manipulation of existing data grows in importance. Finding, interpreting, and manipulating the data already present in MARC bibliographic records to create systematized forms is an invaluable step in moving MARC toward the Linked Data future. The semi-structured data's creation could depend on the original metadata structure and format, or the restrictions of the practices such as before implementation of RDA.

The *OCLC* researchers have been investigating the means by which to find names and their associated role phrases, in order to match those names to authorized forms, and to match role terms and phrases to controlled vocabularies. Approximately 19 million records for musical resources in *WorldCat* were analyzed in 2016. The records were generated during the 45-year history of *WorldCat*, and comprised both musical sound recordings and musical scores. The analysis of these 19 million records determined that they contain approximately 2.5 million names that can be identified as distinct. Roughly 50% of those 2.5 million names have a role designation in the form of a subfield $e (Relator Term), subfield $4 (Relator Code), or both, as well as additional descriptive data that could be mined to further refine the data coded in name fields.

The process can be summarized as:

- from uncontrolled occurrences in notes and/or statements of responsibility in records [e.g., 508 (Creation/Production Credits Note) or 511 (Participant or Performer Note)];
- find named entities and their associated role phrases;
- match them to the corresponding 7XX fields in the same record;
- match names to authorized forms;
- match role terms and phrases to controlled vocabularies.

An encouraging observation is that RDA elements can be incorporated in hybrid records without complete re-cataloging. In this case, it is accomplished by the addition of role designations in access points for music-related metadata.

In matching the identified role phrases to appropriate controlled vocabularies, multiple KOS vocabularies have been used, including the *Library of Congress Subject Headings* (*LCSH*), the *Library of Congress Demographic Group Terms*, and the *Dictionary of Occupational Titles* (*DOT*). Extended work has been conducted on multiple languages for the performer roles, medium of performance terms, associating the name of an instrument with its performer, and more. The experiment also led to the activities of compiling new value vocabularies and mapping to other sources, encompassing the entries in *OCLC*'s *Faceted Application of Subject Terminology* (*FAST*), *LCSH*, *Wikipedia*/*Wikidata*, *Library of Congress Medium of Performance Thesaurus for Music* (*LCMPT*), *Dewey Decimal Classification* (*DDC*), and other multilingual controlled vocabularies.

OCLC's experiments with parsing that data in order to both associate performer names with authority data and identify role terms and phrases with controlled vocabularies have proven fruitful. The potential opportunities for Linked Data across bibliographic and authority data, across vocabularies, and across languages are vast (**Weitz** *et al.*, 2016).

3.2.2. Example: bibliographic metadata in MARC and beyond

Going beyond MARC, the enhancement of library catalogs towards Linked Data will empower library users to discover many more information resources, providing them not just with access to basic descriptive information about works,

but also the context surrounding works' creation, distribution, and use. As pointed out in the *W3C Linked Library Data Incubator Group* (*LLD-XG*)'s final report (*W3C*, 2011), in our current document-based ecosystem, data is always exchanged in the form of entire records, each of which is presumed to be a complete description. In order to integrate library metadata with the Semantic Web, **Dunsire** and **Willer** (2011) laid out detailed examples of how traditional library bibliographic records may be disaggregated into catalog records consisting of RDF triples, as well as what benefits LAMs may receive from including such data. **Alemu** *et al.* (2012) called for a conceptual shift from document-centric to data-centric metadata, moving from MARC-based to RDF-based description, and presented methods to achieve this goal without disrupting current library metadata operations.

In researching metadata's linkability, a team at *Kent State University* in the United States showed how to align metadata structures and properties from diverse communities, specifically how to relate metadata vocabularies familiar in the library community to the unfamiliar resources of the LOD universe (**Gracy**; **Zeng**; **Skirvin,** 2013; **Zeng**; **Gracy**; **Skirvin**, 2013). In this case study, the research team collected, analyzed, and mapped properties used in describing and accessing music recordings, scores, and music-related information across a variety of music genres. The team evaluated 11 music-related Linked Data datasets, 20 collections of digitized music recordings and scores, and 64 representative MARC records for sound recordings and musical scores (and their extended schema.org records). A number of crosswalks were created to align all the data structures, not only by classes and properties, but also indicating the levels of mapping (such as exact match, broad match, narrow match, close match, and no match). The researchers then randomly selected 280 MARC records from each genre to verify the crosswalking results. The process resulted in a unified crosswalk that aligns these properties according to nine common groups of bibliographic data. These include: title information, responsible bodies, physical characteristics, location, subject, description of content, intellectual property, usage, and relation. Many of these properties are usually hidden in the semi-structured data portion of records.

In the areas of performance and recording, the researchers believe that three MARC note (5xx) fields in particular contain valuable data points that could be useful as links to other data sources external to the library catalog, although they are not usable as Linked Data in their current form. Depending on the application and practices of cataloging for sound recordings and musical scores, the linkability could vary, as illustrated by the Figures 11a and 11b.

Other collections with non-MARC metadata reviewed for this study included such information as instrumentation; performance medium; meter; tempo; duration; notes about strains, rendition, phrase structure, assorted musicological details; region and language of music; digital reproduction history; file size; description of the physical container; identifiers such as plate, publisher, label, matrix, and take numbers; and rights information. It is promising that by making bibliographic data shareable, extensible, and reusable, libraries are able to aggregate data based on the pieces/chunks of information they need from a dataset without integrating a whole database or converting full metadata records. They can mash up metadata statements (not whole records) from different namespaces or aggregate data from a variety of resources, based solely on what is needed. (**Gracy**; **Zeng**; **Skirvin**, 2013; **Zeng**; **Gracy**; **Skirvin**, 2013).

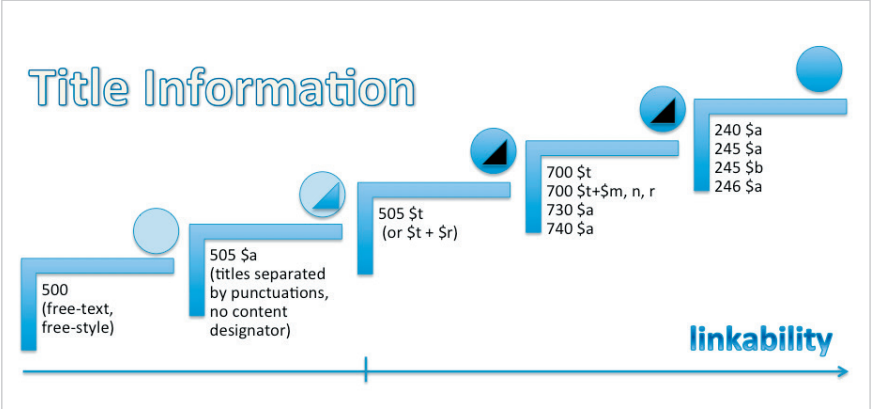


Figure 11a. Illustration of linkability of title information based on the study samples.
Source: **Zeng**; **Gracy**; **Skirvin**, 2013.

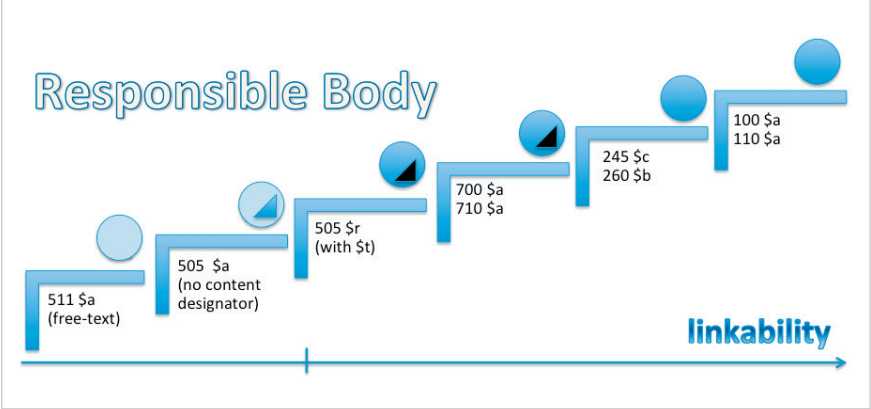


Figure 11b. Illustration of linkability of Responsible Body information based on the study samples.
Source: **Zeng**; **Gracy**; **Skirvin**, 2013.

It should be noted that the cataloging practices and standards have changed during the past five years, hence some of the situations illustrated above might have improved already and the linkability measurement could be updated. Collectively, LAM data as Linked Data is perceived as mainstream since the *W3C*

LLD-XG 2011 reports' release. Today, the number of registered LOD datasets in the old *DataHub* is over 11,000, including 1,100+ named with KOS types (thesaurus, taxonomy, classification, terminology, and ontology) and 1,500+ that corresponded to search by LAMs and special collections, as of October 2018.[14] More research on their role in semantic enrichment of semi-structured will be interesting.

3.2.3. Example: semi-structured data in archive finding aid descriptions

In current archival information systems, useful linkable information is embedded in narrative descriptions known as finding aids. Finding aids are tools that help a user find information in a specific record group, collection, or series of archival materials. Examples of finding aids include published and unpublished inventories, container and folder lists, card catalogs, calendars, indexes, registers, and institutional guides (*National Archives*, 2016). Finding aids can be found in many different formats, ranging from handwritten documents, *Word*, and spreadsheets to HTML and XML. MARC 21 has also been used to encode archival metadata. Increasingly, creating XML-tagged finding aids enabled by *Encoded Archival Description* (EAD) schemas has become a popular practice in the archival community. EAD is a pioneer of metadata schemas developed for the LAM community, with its beta version issued in 1996. EAD has always been among the first to implement the newest markup language standards in the schemas, including from SGML DTD to XML DTD to XML Schema (in Relax NG and in XSD). The semantics/structure and syntax in EAD schemas are integrated as one unit. In other words, EAD as a metadata standard keeps both the specification of the element set and encoding schema in one document.

Finding aids provide contextual information to understand the nature and scope of archival collections. The format of the genre is mainly semi-structured, characterized by large text blocks in which many types of information are intermingled and unlabeled. In the blocks of histories about records creators and the scope notes for collections, many names of persons, organizations, places, and events, as well as topical terms might be mentioned. Apart from full-text searching, the lack of semantic distinction among the different entities and topics hinders efficient and effective information retrieval, which also restricts the ability of information systems to create the links that would gather widely dispersed information about the same person, organization, or thing into one place.

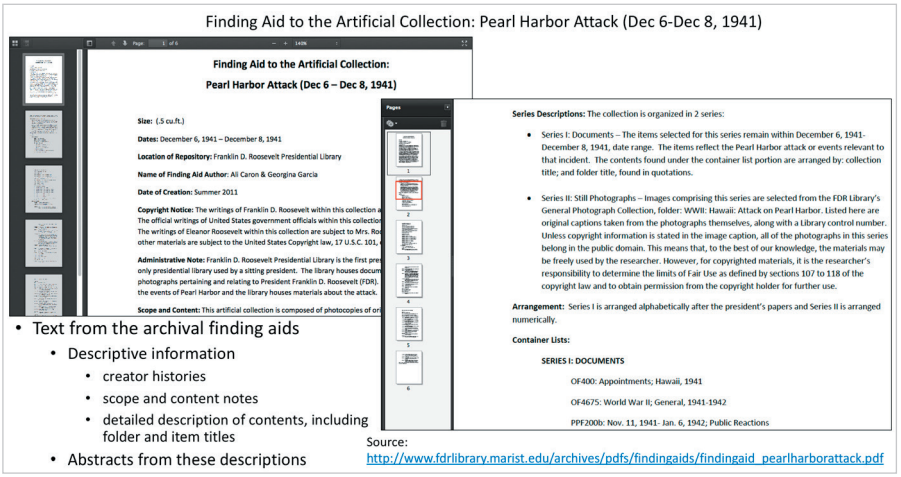


Figure 12a. Portions of a finding aid and explanation of the text used in the semantic analysis process.

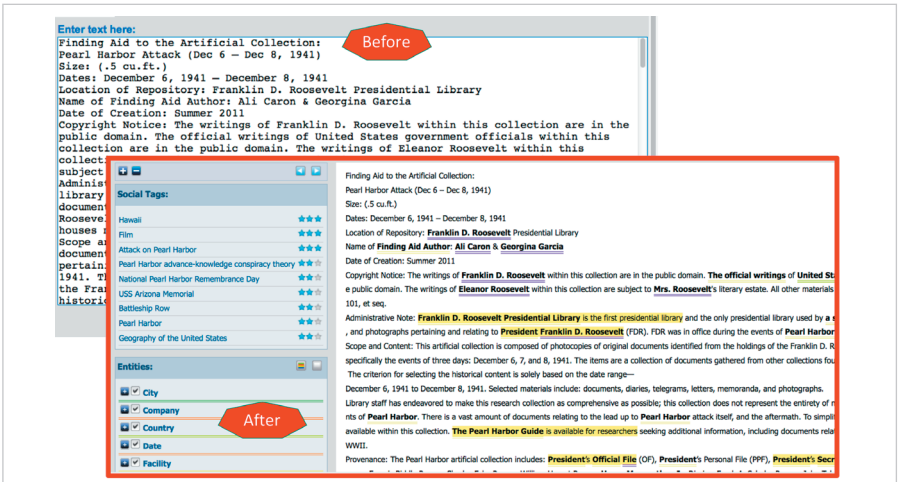
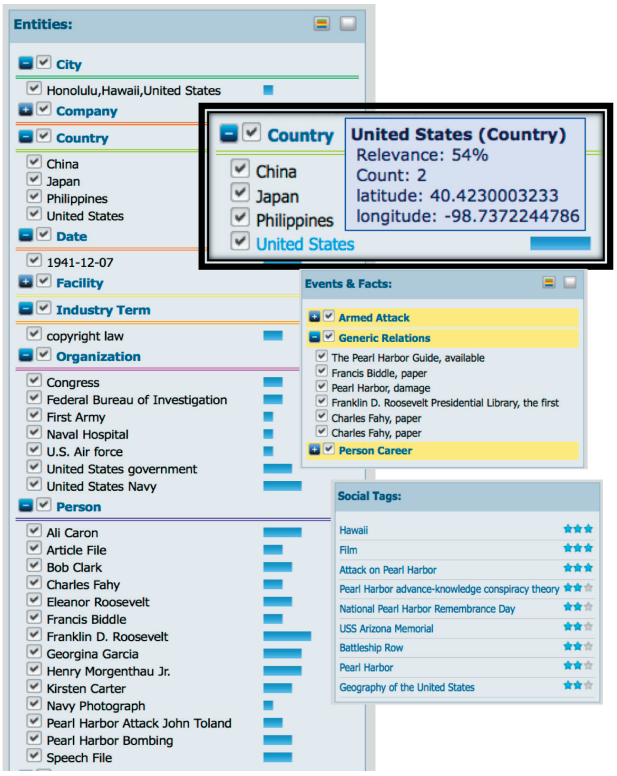


Figure 12b. Example from the semantic analysis results using *OpenCalais* demo tool, indicating the entities and social tags generated.

Inspired by the possible semantic enrichment technology for entity analysis and extracting, as well as the potential of using Linked Data principles to enhance the discoverability and linkability of the rich information in finding aids, Karen Gracy led a research team in conducting an experiment using a sample of 43 archival record groups. The finding aids are from 16 institutions, including university archives, government records archives, and manuscript/special collections repositories in various LAMs. Using the semantic analysis tool *OpenCalais*' free version, descriptive information contained in the archival finding aids (such as creator histories, and scope and content notes sections), as well as abstracts from these descriptions, were used to generate extracted access point candidates (**Gracy**; **Davidson** 2014). [Figure 12a and 12b] *http://www.opencalais.com/opencalais-demo*

The analysis resulted in dozens and, at times, hundreds of potential entities and social tags that could be used to provide additional points of entry to

Findings	Results of semantic analysis by OpenCalais (example)
Entities correctly identified via *Calais* analysis included: - personal names (Person), - corporate names (Company, Facility, Organization), - geographic names (City, Continent, Country, Natural Feature, Province Or State, Region), and - events (Holiday, Political Event). The scores for each identified entity may be used as a valuable clue about the importance of that entity to the overall scope of the archival collection. The "Social Tags," "Industry Term," and "Products" categorizations were the least reliable in terms of accuracy. Analysis of detailed description areas was most likely to lead to incorrect identification of text strings because the descriptions have the physical location information intermingled. Targeted analysis of particular areas of the finding aids may result in better accuracy for topical analysis.	 Figure 13. Findings of the results of automatic semantic analysis using *OpenCalais* demo tool.

these archival records (refer to the Findings in the Figure 13). The dataset of the structured data in RDF/XML format can also be obtained directly from the same output. [Figure 13]

The research team developed a program to allow batch processing. The software automatically obtained the archival records and sent them to the semantic analysis service supported by *OpenCalais*. The output, which was in the JSON format, was then converted directly into a CSV file. The resulting database contained the following fields: Entity-type, Entity-name, Relevance-ratio, and File-source. Using the *OpenRefine* tool, the data were clustered automatically to allow the researchers to clean up the data manually (e.g., merge the synonyms and delete incorrect extractions) and validate the names and topical terms against various controlled vocabularies, such as the *Library of Congress Name Authority File*, *LCSH*, and the *Getty* vocabularies. Figure 14 illustrates this multi-step process. [Figure 14]
http://openrefine.org

The researchers also experimented with other tools to identify named entities and topical terms from finding aids, including *Cogito Intelligence API*, *MachineLinking*, and *Zemanta* (**Gracy**; **Zeng**, 2015).
http://www.intelligenceapi.com
http://www.machinelinking.com/wp
http://www.zemanta.com/api

Figure 15a is an example from the semantic analysis results using *Cogito Intelligence API* demo tool for the selected archival finding aids, showing the initial result preview. The top row circle indicates the options of display, including tagging, categorization, text mining, semantic reasoning, and fact mining, plus by entities such as People, Organizations, and Places (see Figure 15b). The tool also provides Wri-

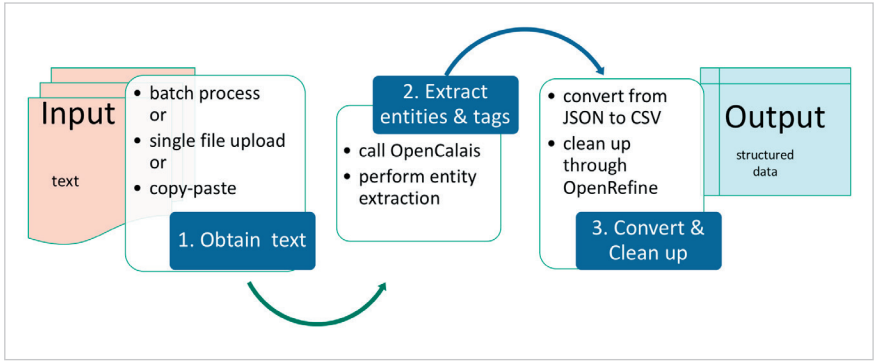


Figure 14. Illustration of the process of entity analysis and extracting

teprint analysis which estimates the readability level of a provided document, collecting and forging a full set of readability indexes as well as grammatical, lexical and semantic analysis scores.

Both *OpenCalais* and *Cogito Intelligence API* are powered by multiple taxonomies and domain ontologies, and feature automatic processes of text mining, categorization, semantic tagging, fact mining, extraction of entities and relationships, and visualization of the entity relationships and geographic locations. In addition to the functions similar to *OpenCalais*, *Cogito Intelligence API* also conducts semantic reasoning. Although neither of these tools have been developed for cultural heritage or humanities domains, meaning that their taxonomies and ontologies are not in these domains, the research results strongly suggest that it would be well worth the effort for institutions to experiment with semantic analysis methods as an initial step to suggest key entities and topics, or as a final check to ensure that important concepts or entities have not been overlooked. For certain types of records, particularly those for which subject indexing is not common, semantic analysis may provide entry points to archival records that were not previously available. Such techniques will enhance subject analysis at the levels of description and identification, but are unlikely to be useful for interpretation of the material. These findings were confirmed through a second case study the research team conducted based on 44 philosophy theses from *KentLINK* and *OhioLINK* (**Zeng**; **Gracy**; **Žumer**, 2014).

3.2.4. Example: semi-structured data in item-centered information cluster webpages

Comparable to the processes for enhancing the semi-structured data in archival finding aids records, many information cluster webpages also show the infinite potential of semantic enrichment for expanded access points through the semi-structured portions. The uniqueness of the examples to be discussed next demonstrates that, ultimately, the enrichment process can be implemented case-by-case, collectively or independently, with or without significant project funding.

By testing various types of descriptive information clusters carried by webpages, it is clear that the semi-structured data that have been created for information access and exposure can be semantically enriched effectively. Examples of item-centered information cluster webpages that were tested by the author include metadata descriptions about photograph special collections, table-of-contents included in metadata records, back-of-book indexes for transcribed oral history materials, and captions and other descriptive information for cultural objects presented on their digital representation webpages. They demonstrate a limited number of types, while denoting that similar results can be obtained by many other types of semi-structured data resources.

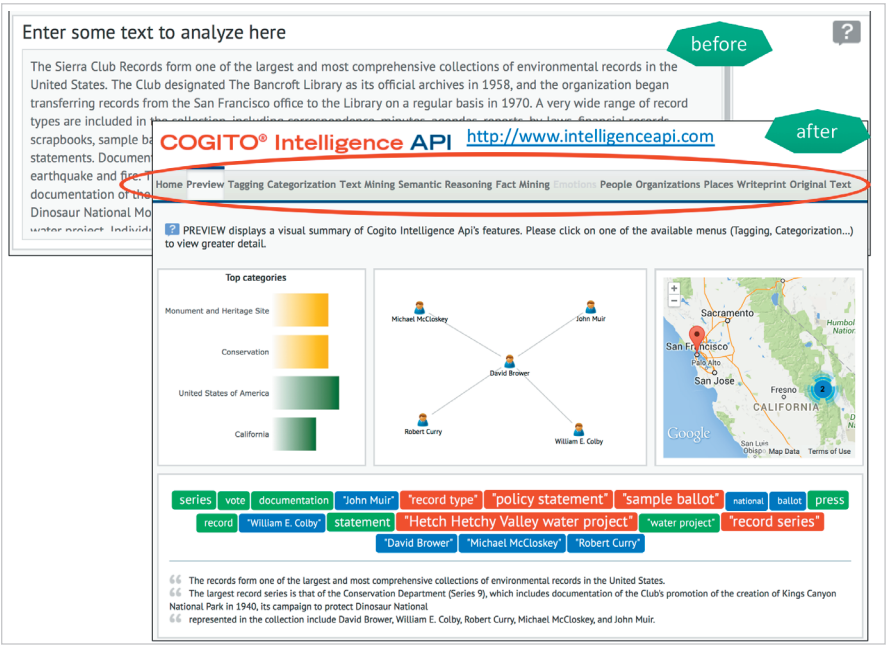


Figure 15a. Example from the semantic analysis results using *Cogito Intelligence API* demo tool.

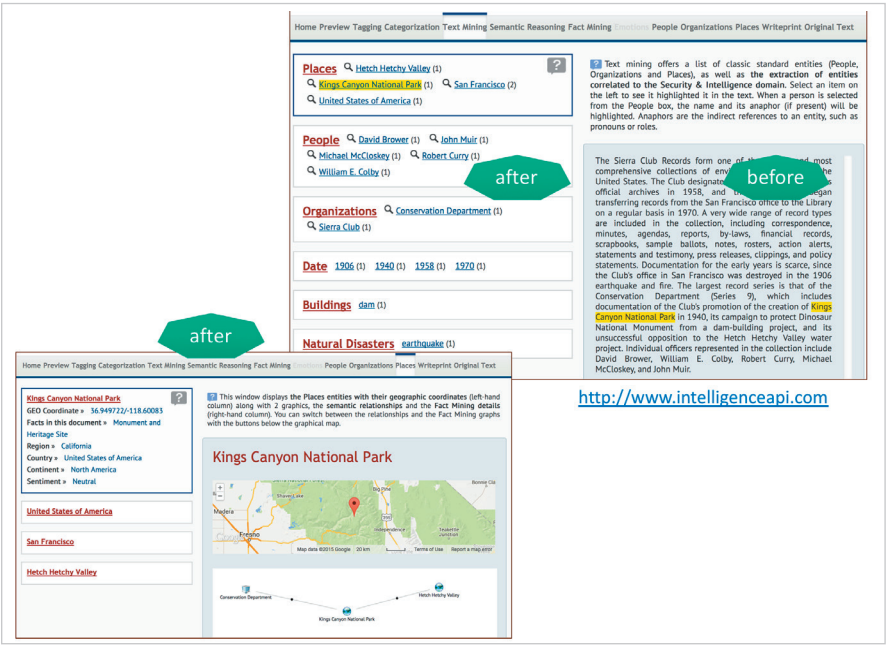


Figure 15b. Example from the semantic analysis results using *Cogito Intelligence API* demo tool, indicating the text mining and named entity (places) generated.

The results of newly generated structured data can be embedded in individual webpages or integrated into databases.

Metadata and curator's notes for a special collection

Named entities can be extracted from free-text segments in metadata descriptions about a special collection. In this example, entity extraction generated expanded access points based on place names of a special collection of thousands of historical photographs. [Figure 16a]

Table of contents included in metadata records or linked by metadata records

Table of contents (TOC) included in or linked by a metadata record of an edited book could be sources for entity extraction. In this example, shown in the Figure 16b, the TOC as entered in the original record (left in Figure) or generated by machine (center in Figure) provides expanded access to the authors and themes of this item (right in Figure). [Figure 16b]

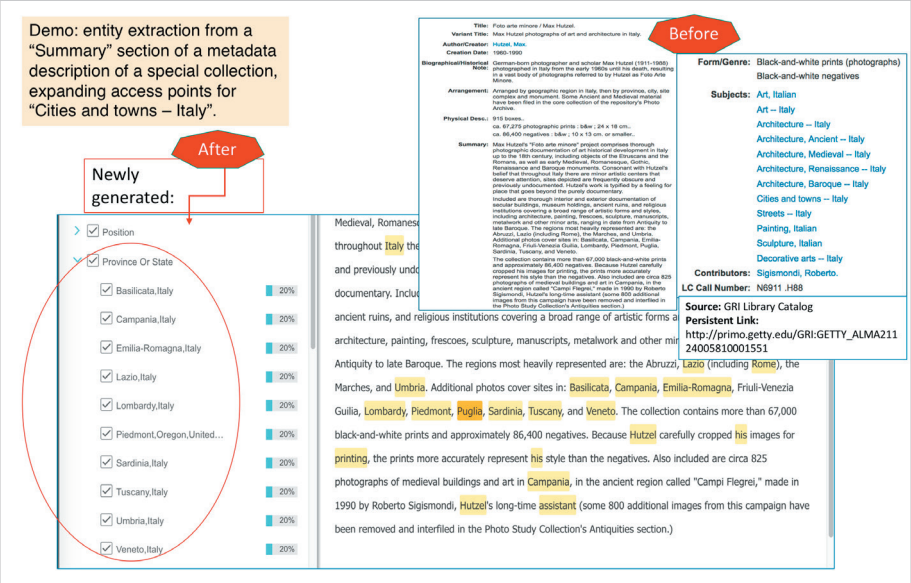


Figure 16a. Example of entity extraction (place) from a Summary section of a metadata description.

Back-of-book indexes for transcribed oral history materials

Similar to the TOC, back-of-the-book style indexes for oral history materials are precious sources for generating new access points.

Due to widespread digitization efforts over the last 20 years, many of the oral history materials hosted by LAMs have transcripts available in PDF or other digital formats, according to the copyright and privacy conditions. The oral history transcripts files might be managed at the collection level only, or they may be indexed using back-of-the-book style and kept together with PDF files that are downloadable. The indexes are usually established with high quality and involve collaborations between different units and institutions. At the highest level these indexes are processed for text-based searching. There is great potential for applying semantic enrichment method to these information products.

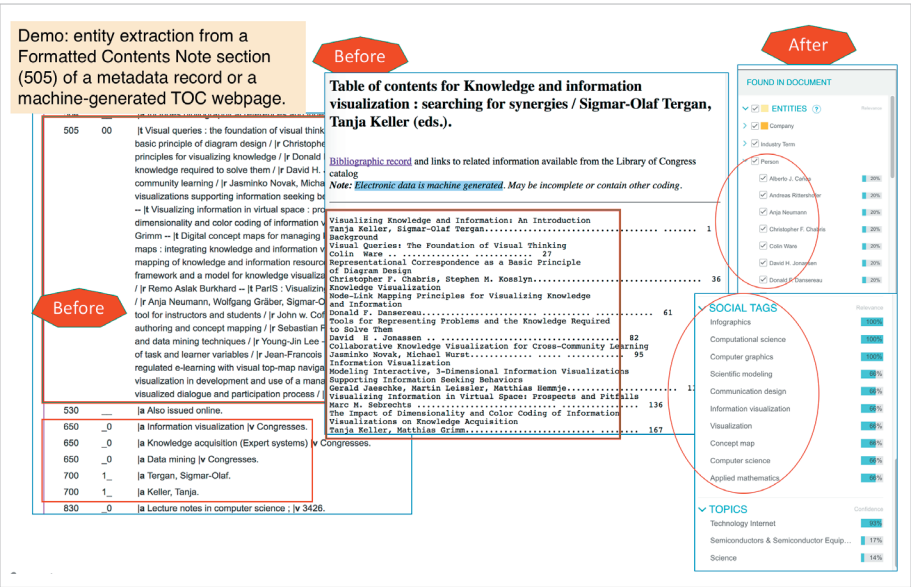


Figure 16b. Example of entity extraction from TOC provided by the metadata record of a publication.

Captions and other descriptive information for cultural objects presented on their digital representation webpages

Captions and other informative descriptions for cultural objects featured on the individual webpages of cultural institutions could also benefit from entity extraction. In this example, the text (see <Before> in the Figure 17) about four ancients and their favorites depicted on a vase made during the Yuan Dynasty (1271-1368) are recognized by machines (see <After> in Figure 17) through entity extraction (using *Boson*, *OpenCalais*, and *Cogito*). [Figure 17]
https://bosonnlp.com/demo

Although this test was performed using free tools and resulted in different degrees of completeness, the potential of generating information access and discovery based on the entities and keywords is promising. Such benefits might be

especially meaningful to institutions with relatively small digitally represented collections, including those in non-English-speaking locations or less digital-driven places.

3.2.5. Discussion

An important feature of semi-structured data resources that should be recognized actually resides in their nature of being the products of information processing. These semi-structured data represent the accumulated time, knowledge, and experience of the creators who generated these metadata through formal workflow which conforms to professional standards and best practices. With semantic enrichment processes, the data values in semi-structured data are contextualized through the metadata elements/fields; hence, the function and meaning are clearly implied. By parsing these data through advanced information technologies, these LAM data are dramatically enriched and are converted into new access points.

The *OCLC*'s experiments on bibliographic records revealed potential opportunities for Linked Data across bibliographic and authority data, across vocabularies, and across languages. Other pilot studies shared in this section also prove that, by making bibliographic data shareable, extensible, and reusable, LAMs will be able to aggregate data based on the pieces/chunks of information they need from a dataset without integrating a whole database or converting full metadata records.

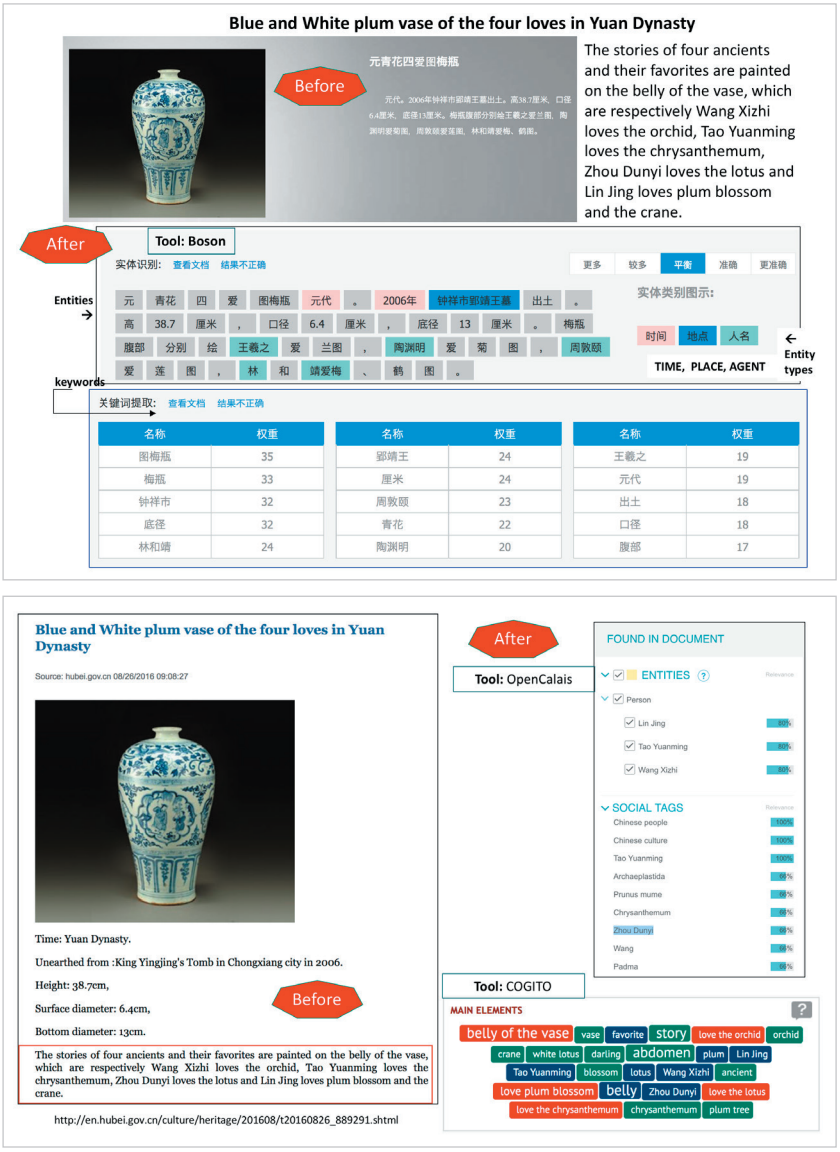


Figure 17. Example of entity extraction from captions and other descriptive information for a cultural object using three different semantic analysis tools.

The examples presented in this semi-structured data enrichment section reveal that, ultimately, additional useful data can be derived from large digital collections as well as from individual item-centered information clusters. These activities can be managed case-by-case, from the top-down or the bottom-up, collectively or independently, with or without significant project funding.

Semantic analytics, one of the advanced semantic enrichment methods, has been used for analyzing, searching, and presenting information by using explicit semantic relationships between known entities. It is a major method used in the semi-structured data processing presented above. The tools used in the experiments discussed in this section, such as *OpenCalais* and *Cogito Intelligence API*, are powered by multiple taxonomies and domain ontologies, and benefit from machine learning and other new artificial intelligence (AI) technologies, far beyond normal natural language processing. The APIs classify entities using different taxonomies and disambiguate them with different knowledge bases. Processes include recognizing named entities mentioned in text, assigning them as pre-defined types, and linking them with their matching entities in a knowledge base. "Entity" has been a hot keyword ubiquitous in semantic technology related conferences, such as the most recent 2018 *International Semantic Web Conference*, where "entity"-related research tracks range from entity extraction, annotation, recognition, disambiguation, to relation linking and embedding, while the entity of interest could be varying.
https://link.springer.com/book/10.1007%2F978-3-030-00671-6

Countless semantic analysis and machine learning experiments and tools have been reported and can be found in literature. Taking entity disambiguation as an example, numerous algorithms can be applied to measure text string si-

milarity, semantic or structural similarity, contextual similarity, and commonness. Similar to what *Cogito* has also included as one of its semantic analysis results' delivery, Knowledge Graphs (KG) are widely used abstractions in representing entity-centric knowledge. On top of rule-based systems, embedding-based systems for Knowledge Graph completion has become a dominating focus in research and development during recent years, also revealed by those reported at the *2018 International Semantic Web Conference*. The role of LAM data created for cultural objects will be both the consumer and the contributor of semantic technologies and the web of data.

> "An important feature of semi-structured data resources actually resides in their nature of being the products of information processing. These semi-structured data represent the accumulated time, knowledge, and experience of the creators who generated these metadata through formal workflow which conforms to professional standards and best practices. Their semantic enrichment activities can be managed case-by-case, from the top-down or the bottom-up, collectively or independently, with or without significant project funding"

3.3. New structured data generated from unstructured data, supporting knowledge discovery

The semantic enrichment approaches and issues discussed in this section will be about information discovery and re-discovery from unstructured data. In data resources that are served through LAMs, unstructured data are usually available in the largest quantity in comparison with the structured data; have the most diversity in type, nature, and quality; and are the most challenging to process. In LAMs, they can be found in documents and other information-bearing objects (textual or non-textual, digitized or non-digitized), in all kinds of formats. This section will focus on certain types of LAM data hosted at retrospective resource warehouses. The mixed or heterogeneous contents from distributed data providers will be discussed in the last section of this chapter (Section 4).

In terms of text-based information processing and retrieval, for a long time, text-matching (especially word matching) had been the primary way to do full-text searching. Only recently, content-based searching (according to semantic meanings) has been implemented and used on a large scale, as symbolized by Knowledge Graphs. To advance discovery from text-based resources, instead of relying on full-text searching, methods such as semantic-based analysis, extraction, mining, and tagging are used to improve the information discovery from unstructured data. [Refer to the examples introduced in the previous section (Section 2) on content-based new approaches and tools].

For non-text-based resources, semantic annotation and ontology-based knowledge bases have revealed great potential for digging into data and supporting digital humanities research. In general, "annotating" is the act of expressing knowledge about a resource, attaching names, attributes, comments, descriptions, etc., to a document or to a selected part in a resource. Annotating provides additional information (metadata) about an existing resource. "Semantic annotation" goes one level deeper: it enriches the unstructured data with a context that is further linked to the structured knowledge of a domain.

Tools such as *Oasis* (*Open Annotation Semantic Imaging System*), *Pundit*, *Mirador*, *Brat Rapid Annotation Tool*, *Recogito*, and *MapHub* enable users not only to comment, bookmark, and tag, but also to create semantically structured data while annotating. With effective application of *W3C* standards, these annotation results are semantically marked up and expressed with typed relationships, teaching a computer how data items are related and how these relations can be evaluated automatically. In computerized systems, further semantic enrichment is based on the annotations. From there, the decision can be made according to logic, to come up with knowledge insertion. It also helps to link other same entities or related entities, which would be very useful for semantic analysis, as illustrated by **Daniel Mayer**'s video (2011), *Mainstream Semantic Enrichment*.
https://www.synaptica.com/image-annotation-indexing
http://thepund.it
http://projectmirador.org
http://brat.nlplab.org
https://recogito.pelagios.org
http://maphub.github.io

The following subsections introduce cases representing semantic enrichment on various types of LAM data, including oral history transcripts, OCR-ed materials, digital counterparts of cultural objects, maps, and images. These cases and research experiments are among the pioneering works related to Linked Open Data and digital humanities. Inspiring news of similar projects are reported more and more worldwide.

3.3.1. Example: oral history transcripts

The last five years have witnessed the widespread use of "Linked xyz" titles in cultural heritage discovery in digital environments, after the *W3C Library Linked Data Incubator Group Final Report* was released (*W3C*, 2011). Among these titles are highly recognized examples acknowledged by the LODLAM community such as: projects (*Linked Jazz*, *Linked Heritage*, *Linked Taiwan Artists*, *Linked Maps*); models (e.g., the *Linked Art Data Model* from the *Linked Art Community*);

and ontologies (e.g., a chain of *Linked Building Data* ontologies by the *Linked Building Data Community Group*).
https://linkedjazz.org
http://www.linkedheritage.eu
http://linkedart.ascdc.tw
http://usc-isi-i2.github.io/linked-maps
https://linked.art/index.html
https://www.w3.org/community/lbd

The *Pratt Institute*'s *Linked Jazz* project can be regarded as a forerunner in the field, presented as a finalist in the *2013 LODLAM Challenge* competition.
http://summit2013.lodlam.net

The project draws on jazz history materials in digital format to uncover meaningful connections between documents and data related to the personal and professional lives of jazz artists, and expose relationships between musicians that reveal their community network (**Pattuelli**, 2012). The 50+ interview transcripts were from various resources (the *Rutgers Institute for Jazz Studies Archives*, *Smithsonian Jazz Oral Histories*, the *Hamilton College Jazz Archive*, *UCLA*'s *Central Avenue Sounds series*, and the *University of Michigan*'s *Nathaniel C. Standifer Video Archive of Oral History*).[15] The documents were in PDF and text format, ranging from 12 to 187 pages in length.

The *Linked Jazz* project provides clear roadmaps for others to follow:

- digging into unstructured data, applying semantic analysis and annotation;
- mashing-up using *DBpedia*;
- establishing name authorities based on *VIAF* and *DBpedia*;
- developing an ontology for relationships (e.g., *knows*, *mentorOf*, *isInfluencedBy*, *collaboratedWith*) after consulting *Friend of A Friend (FOAF) ontology* and *Music Ontology*;
- crowdsourcing on assigning more granular terms to describe the relationship between an interviewee and the person mentioned (for the *Linked Jazz 52nd Street* sub-project); and
- visualizing the networks with images, videos, and short biographies of jazz musicians within the networks. [Figure 18]

The *Linked Jazz* project sits at the intersection of three important domains:

(1) LOD concepts and semantic technologies to be used in LAMs;

(2) the roles and contributions of knowledge organization methods in LOD-enabled products; and

(3) the applicable areas beyond bibliographic data and conventional resources' management, i.e., the unstructured data that LAMs have managed for years but could be better used in discovery.

It utilized solid research methodologies, implemented international standards, and applied innovative technologies to the digitized oral history transcripts from jazz archives for the discovery, visualization, and use of primary sources.

The efforts of digging into these oral history transcripts have revealed new discoveries to the world that are highly regarded by the jazz education community. Extended research projects are adding facets representing various professional

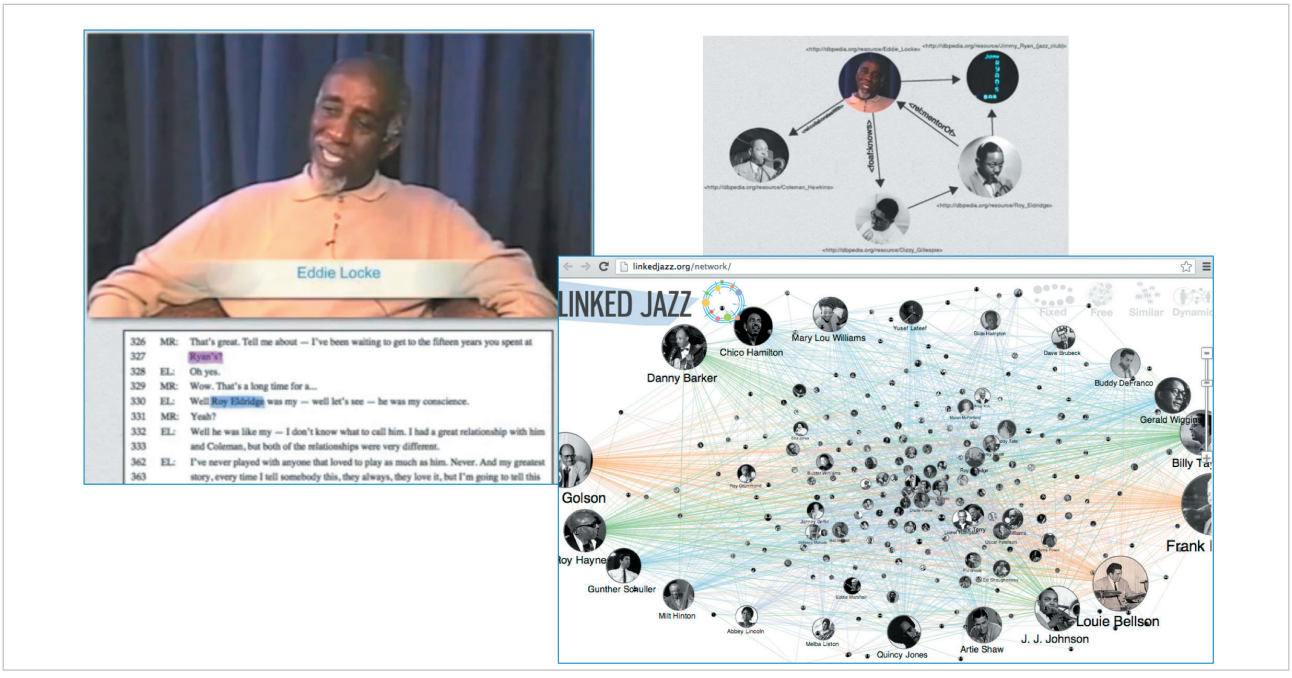


Figure 18. Demo of *Linked Jazz* project outcomes. *http://linkedjazz.org*

and social aspects of the jazz community, and expanding data sources to include other document types and music-related datasets (**Thorsen**; **Pattuelli**, 2016; **Pattuelli**; **Hwang**; **Miller**, 2016).

3.3.2. Example: OCR-ed documents

Analogous to the transcribed documents are a large number of historical documents that have been obtained through Optical Character Recognition (OCR), which involves translating the documents into machine processable text.

Before any data mining, entity extracting, and analysis can be done for OCR-ed documents, one has to face the challenges for character recognition, due to various reasons such as font disparities across different materials, lack of orthographic standards where the same word might be spelled differently, material quality, and the unavailability of lexicons of known historical spelling variants (**Mutuvi** *et al*., 2018; **Lin** *et al.*, 2012). Yet, such aspects and their impacts on semantic analysis have been less explored so far in comparison with other topics on digital libraries (**Hinze** *et al.*, 2015; **Bainbridge** *et al.*, 2016). The two research projects introduced next represent the efforts to fill such a gap.

Approaches for OCR post-processing can be seen through three main categories: manual error correction, dictionary-based error correction, and context-based error correction (**Nguyen** *et al.,* 2018). Obtaining thematic patterns from large unstructured collections of text by grouping documents into coherent topics is another major approach, among which common topic modeling techniques are the *Latent Dirichlet Allocation* (LDA) and the *Non-negative Matrix factorization* (NMF). **Mutuvi** *et al.*'s research examines the effect of noise on unsupervised topic modeling algorithms, through comparison of performance of both the LDA and NMF topic models in the presence of OCR errors. The research is supported by the *NewsEye* project, which is funded by the European Union's *Horizon 2020* research and innovation program. *NewsEye* is aimed at tens of millions of newspaper pages from European libraries that have been digitized and made available online during the last decade.
https://www.newseye.eu

In its poster reporting the inner testing of newspaper datasets from three partner libraries, the processes are laid out as:

- text recognition & article separation;
- semantic text enrichment;
- dynamic text analysis;
- personal research assistant; and
- user interface,

where the "semantic text enrichment" processes include: named entity recognition, stance detection, novelty detection, and event detection.
https://www.newseye.eu/fileadmin/user_upload/Poster_white.png

The research team of the *Capisco* project at *University of Waikato*, New Zealand, conducted research to address OCR-ed documents (**Bainbridge** *et al.*, 2016). The rationale is that lexicographic search in large collections, such as the *HathiTrust*'s 13 million volumes with 4.6 billion pages, often returns large sets of unrelated documents (due to homographs —same spelling, different meaning-being included), while relevant sources may remain undetected unless the right keyword is found. The problem is exacerbated in documents that have been obtained through OCR, as recognition errors may lead to misidentification of terms, which are then either mistakenly included or omitted from the search results. Meanwhile, the research team is motivated to avoid major changes to the document retrieval mechanism and indexing strategy, query processing, and interface to which users and technical teams are accustomed.

The *Capisco* system developed by the research team focuses on semantic indexing and search. It uses a knowledge base containing information about concepts in context, initially created by mining *Wikipedia* and potentially further enriched by domain experts. Each concept is identified by an ID, and also carries a human-readable concept label. There are several key steps:

- Concept labels are derived from *Wikipedia* article titles.
- Synonymous terms for a concept are stored with reference to the context in which they appear.
- The context of a term refers to the main area in which this term is used for this concept.
- Because contexts are also concepts, the knowledge base forms an interlinked Concepts in Context (CiC) network.

The knowledge base is used to disambiguate a term (i.e., identifying its semantic concept) and identify potentially matching concepts, which would lead to the identification of significant topics within a document. For each context, an index entry is created with references to the pages on which the term appears.

The *Capisco* project team reported an analytical approach to explore five strategies for low-cost semantic enhancement to large digital collection's metadata and indexing, especially the OCRed historical materials, showing through examples how using semantic concepts can help identify OCR errors. In testing the results, five approaches were used:

- Approach 1. Concept labels added to metadata
- Approach 2. Concepts and synonyms added to metadata
- Approach 3. Concept labels indexed at page-level

- Approach 4. Concept label and synonyms indexed
- Approach 5. Concepts and synonyms added at page metadata (only possible for digital library implementations that support page-level metadata fields)

Three of these approaches address adding information about semantic concepts to the metadata (1, 2, 5) and two others concern adding information about semantic concepts to the full-text index (3 and 4). A case study of four documents showed the differences in result sets for lexical search, semantic search, and each of the five approaches. The study also established four collections (with nearly five thousand pages and thousands of tokens and concepts indexed) in which the performance implications of enriching the full-text index with concept labels was experimentally determined. Based on semantic enrichment, search is now possible via both the simple interface and the advanced interface for filtering and search in metadata-based enhancements. The researchers concluded that even though the work test focused on simply enhancing the lexical search capabilities of traditional digital libraries, the most powerful solution would be a combination between lexical-based search and semantic search as offered in *Capisco* (**Bainbridge** *et al.*, 2016).

3.3.3. Example: historical maps

Maphub, another pioneer project, is an online application for exploring and annotating digitized, high-resolution historic maps, developed at *Cornell University*'s *Department of Information Science*. All user-contributed annotations are shared via the *Maphub Open Annotation API*. The first demo was bootstrapped with approximately 6,000 public domain maps taken from the *Library of Congress Historic Map Division*.
http://maphub.github.io
http://maphub.github.io/api

The *Maphub Open Annotation API* follows the *W3C Open Annotation* specification and uses *Apache Solr* for map full-text search.
https://www.w3.org/TR/annotation-model
http://lucene.apache.org/solr

Annotations can easily be added by creating overlays on top of map images. When a user opens a map to annotate zoomable raster images, *Maphub* suggests potentially relevant *Wikipedia* tags. The semantic tag enrichments are retrieved from *DBpedia*. By aligning with *DBpedia*, it is possible to exploit those connections to enrich annotations and their tags with additional information, such as the ability to search for a map by its content and not its title, and translations of terms in other languages.

The distinctive feature of this project is that, after adding at least three control points to a map, it is possible to calculate real world locations for any point on the map. This allows users to create different views through *Google Maps*. The historical map can also be laid over the *Google Earth* map, creating an overlay of the historic map onto its current day location. The *Google Earth* file can also be downloaded to a computer for later viewing.[16]

3.3.4. Example: images of cultural heritages

Over the last two decades, the web welcomed all kinds of new digital collections, domain-specific information resource portals, online exhibitions, and other products that expose and provide access to the not-born-digital resources hosted in LAMs. Digital images are a container for much of the information content in web-based delivery of images, books, newspapers, manuscripts, maps, scrolls, single sheet collections, and archival materials. Many of them are not good candidates for OCR, and therefore usually remain as digital images. In the digital age, access to these image-based resources is fundamental to research, scholarship and the transmission of cultural knowledge. On the other hand, cultural heritages objects, though steadily being digitized, are still difficult to find, reuse, cite, exchange, and compare.

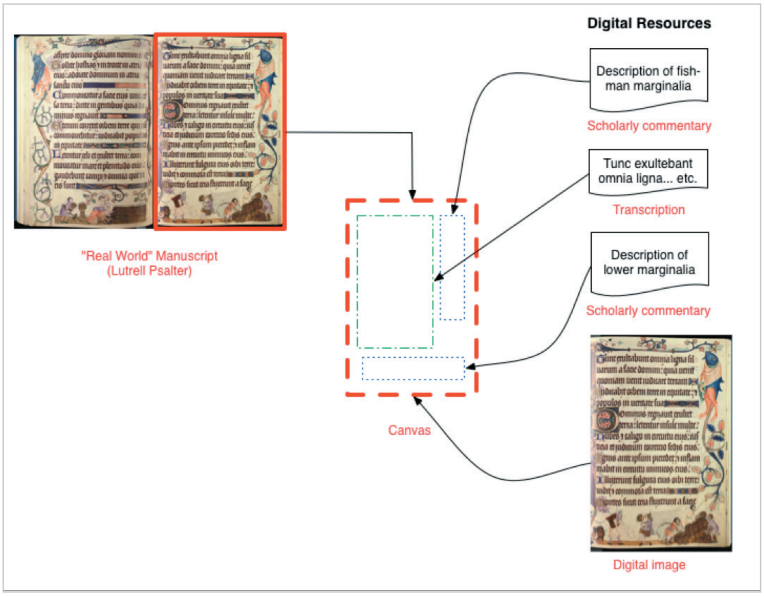


Figure 19. *Shared Canvas Data Model*.
Source: **Albritton**, 2013. Slide 12.

Image exchange APIs

As pointed out by the community that developed *IIIF* (*International Image Interoperability Framework*) APIs, much of the Internet's image-based resources are locked up in silos, with access restricted to bespoke, locally built applications.[17] API is the

abbreviation for "application programming interface", and is compiled with a set of routines, protocols, and tools for building software applications. To address the issue of interoperability and provide access to high quality image resources across all kinds of digital collections, four *IIIF* APIs have been developed, tested, and revised during the last five years and are already supported by major software for image servers, image clients, and image apps. They are: Image API, Presentation API, Content Search API, and Authentication API. An API for Audio/Visual materials is in progress.

One of the important fundamental concepts of *IIIF* is the standardized URI patterns. For example, the Image API specifies a web service that returns an image in response to a standard HTTP or HTTPS request. The URI can specify the region, size, rotation, quality characteristics, and format of the requested image. A complication for Image APIs is the varying quality, size, and format of the source images involved in the viewing, zooming, searching, annotating, exchange, etc. The *IIIF* Image API URI for requesting an image *must* conform to the following URI Template (**Appleby** *et al.*, 2012a):
https://iiif.io/api/image/2.1/#image-request-uri-syntax

{scheme}://{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}

For example:

http://www.example.org/image-service/abcd1234/full/full/0/default.jpg

Using the supported tools, annotations by experts or other contributors can be added to the canvas. Annotations, annotation lists, and content all have unique identifiers that can be processed by machines without confusion. [Figure 19]

Different resource types defined by the framework follow the recommended URI patterns (**Appleby** *et al.*, 2012b): [Figure 20]

Hundreds of successful cases of digital collections have been benefited from *IIIF* APIs. One example is *The National Library of Wales - Welsh Newspapers Online* which currently has 1.1 million *IIIF* images. Implementations of *IIIF* at many LAMs can be found via the demo links from the *IIIF* website as well as from the *IIIF* conferences held worldwide.
http://newspapers.library.wales

Deep image annotation

Deep image annotation (DIA) is becoming an accepted new concept in treating images for semantic enrichment, aiming to improve the efficiency of image retrieval and access, enhance users' understanding of an image, and support automatic image resources integration and knowledge discovery.

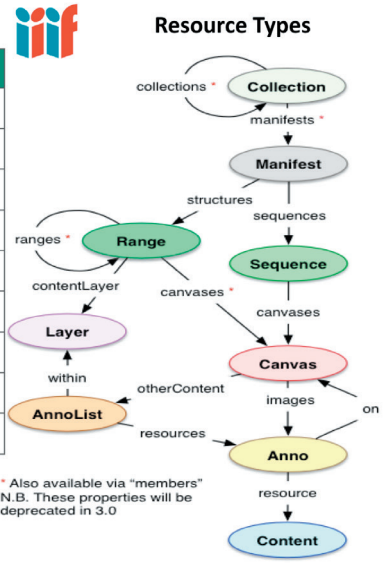
It is widely recognized that a knowledge gap exists between access to knowledge in textual content versus image-based content. A vast amount of information resides inside images, photographs, paintings, diagrams and drawings which can be seen but not searched. It is waiting to be discovered, but has been inaccessible to traditional query methods. In addressing knowledge organization and discovery objectives for images, **David Clarke** (2015) provided a line of objectives:

1. Every image should have a URI and an extensible set of metadata.
2. Visual features within images should have URIs and an extensible set of metadata.

Recommended URI Patterns

Resource	URI Pattern
Collection	{scheme}://{host}/{prefix}/collection/{name}
Manifest	{scheme}://{host}/{prefix}/{identifier}/manifest
Sequence	{scheme}://{host}/{prefix}/{identifier}/sequence/{name}
Canvas	{scheme}://{host}/{prefix}/{identifier}/canvas/{name}
Annotation	{scheme}://{host}/{prefix}/{identifier}/annotation/{name}
AnnotationList	{scheme}://{host}/{prefix}/{identifier}/list/{name}
Range	{scheme}://{host}/{prefix}/{identifier}/range/{name}
Layer	{scheme}://{host}/{prefix}/{identifier}/layer/{name}
Content	{scheme}://{host}/{prefix}/{identifier}/res/{name}.{format}

Resource Types



http://iiif.io/api/presentation/2.1/

Figure 20. Recommended URI Patterns for *IIIF* resource types.
Source: Compiled based on the images from *IIIF* Implementation API specification.

3. Visual features should be identifiable in-line with the image, i.e., by using point-to- bound-area markers.
4. Visual features should be presentable as alphabetical lists.
5. Visual features should be presentable as ordered lists or hierarchical structures.
6. Images and visual features should be indexable using controlled vocabularies.
7. All image, annotation, and indexing metadata should be searchable.
8. Users should be able to pan-and-zoom to specific parts of an image from search results or browse lists.
9. Users should be free to pan-and-zoom anywhere on an image and discover the visual features in view and their related concepts.
10. The system should support the discovery of images and parts of an image that are conceptually related to any other image or part of an image. (**Clarke**, 2015, slide 12).

The *IIIF* APIs have enabled some of these goals to be reached. Other objectives spelled out in this list have been realized by tools like *Oasis*, *Mirador*, *Luna*, *Universal Viewer*, and others.

The following example demonstrates the deep semantic annotating results using *Oasis* for the *Dunhuang Mogao Caves*' "*Nine-colored deer*" painting (**Wang**; **Liu**; **Xia**, 2017). *Oasis* enables the visual features of an image to be individually identified and expressed as Linked Data URIs. These features can then be semantically indexed to internally or externally curated KOS vocabularies. The annotation documented on the left side of the figure represents hierarchical and ordered list structure, expressed automatically in SKOS properties to form a new LOD KOS vocabulary. [Figure 21a]

The narrative story painting of the *Nine-colored deer Jataka*, on the wall of the *Mogao Grottoes Cave* 257,[18] consists of eight episodes, in which the *Nine-colored deer* and other major characters occur multiple times. It reminds us of the value of semantic image annotating (with contextual information) which generates machine-processable data (not just machine-readable data). [Figure 21b]

Existing methods for indexing features and themes appearing in images of works are usually found at the metadata level. The existing metadata standards and cataloging rules provide detailed guidance about external characteristics of images. On the other hand, content-based semantic analysis and annotation models and practices have not been standardized; they are usually project-based, and are therefore less useful in discovering images and parts of an image that are conceptually related to any other image or part of an image. The impact goes to the granular resource aggregation and knowledge discovery of cultural heritages. For this reason, researchers of the Mogao Caves at *Wuhan University* in China established a workflow for image representation and annotation related tasks, with three semantic annotation models: (1) the macroscopic concept model, (2) the information hierarchy model, and (3) the structured image annotation model. The testing results prove that semantic enrich-

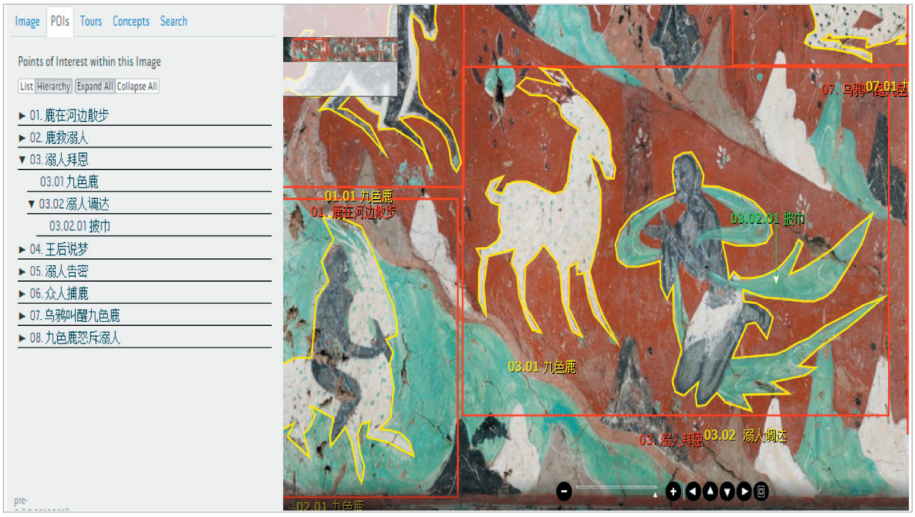


Figure 21a. A section of the semantic deep image annotation result.
Source: **Wang**, **Liu**, and **Xia** (2017).

Figure 21b. Image of the *Nine-colored deer Jataka*.
Source: *Dunhuang Academy*. *Digital Dunhuang*. Mogao Grottoes Cave 257, West Wall.
https://www.e-dunhuang.com/cave/10.0001/0001.0001.0257

ment can be significant in both the development of image information organization methodology and digital humanities research (**Wang**; **Liu**; **Xia**, 2017).

3.3.5. Example: cultural objects

Memory institutions and information services are increasingly embracing the new information technologies in order to meet the needs of the changing digital age. The challenges of three-dimensional objects are unique, in comparison with text-based unstructured data and two-dimensional materials. Innovation is needed because "data are for discovery and inspiration, not just management." (*Opencontext.org*)
https://opencontext.org

Online Coins of the Roman Empire (*OCRE*), was initiated as a joint project of the *American Numismatic Society* and the *Institute for the Study of the Ancient World* at *New York University*.
http://numismatics.org/ocre

This revolutionary new service is designed to help in the identification, cataloging, and research of the rich and varied coinage of the Roman Empire. The original goal of *OCRE* was to make available a digital corpus of all published Roman Imperial coin types. At the same time, the project aimed to expand the ability of all external contributors interested in linking any collection-based online catalog (**Reinhard** *et al.*, 2017). Since the first edition of *OCRE* launched in late 2012, the project has grown tremendously with international collaborations. All coin types from Augustus to Zeno (representing five

Built with Linked Data technology, *OCRE* becomes a knowledge base, much more than a traditional website.

While the searching, browsing, querying and visualization are supported by the Sparql queries, triple stores, and various apps, its ontology-based design make it an easy-to-use digital corpus, with downloadable catalog entries, incorporating over 43,000 types of coins. Users are provided with the multiple browsing, searching, and refining options.

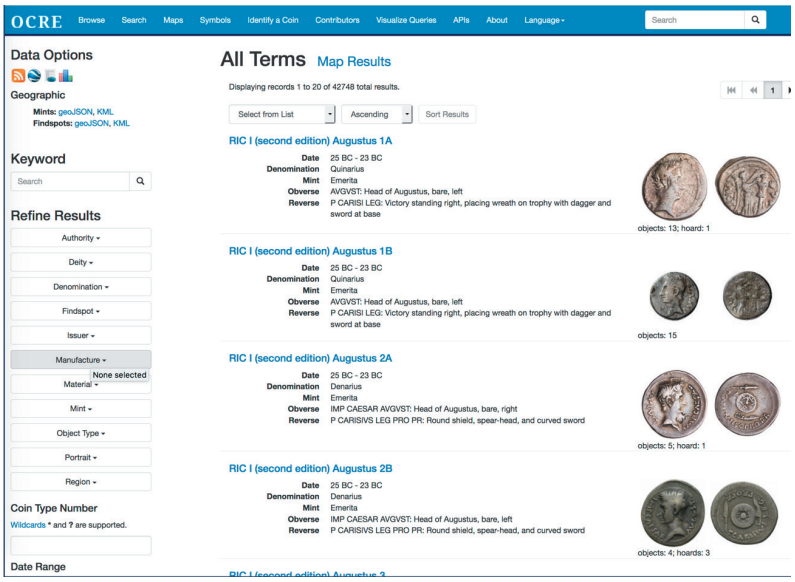


Figure 22a. *OCRE* browse page.
http://numismatics.org/ocre/results

OCRE offers detailed structured data about objects' typological descriptions, geo maps, and examples, as well as Quantitative Analysis options.

In this example, in the Quantitative Analysis section, data about the average measurements for this coin type are provided. A user can also select measurements (by axis, diameter, or wright), choose a chart type (bar or column), and request results of comparison according to the selected categories (Denomination, Mint, Region, Manufacture, Material, Authority, Portrait, and Deity).

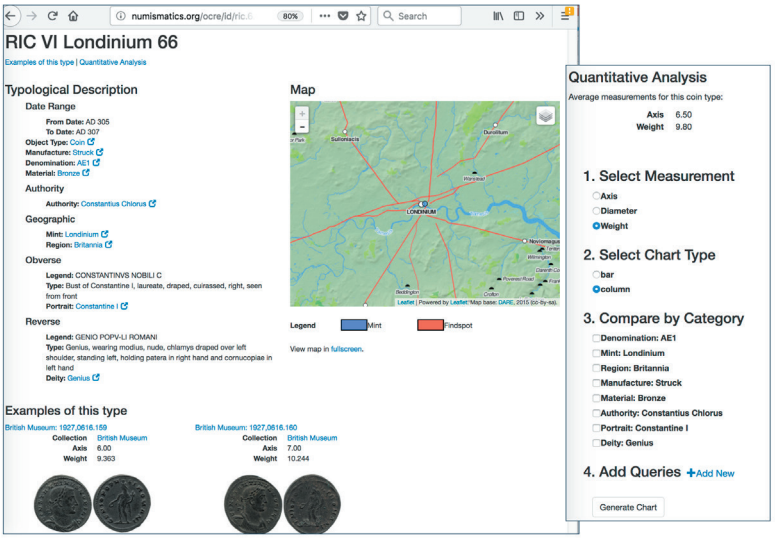


Figure 22b. *OCRE* individual coin type page.
http://numismatics.org/ocre/id/ric.6.lon.66

centuries of Roman imperial numismatics) have been published. *OCRE* incorporated more than 107,000 physical coins related to these coin types from 21 different datasets. These datasets originate from large collections as well as smaller civic or university museums, archaeological databases, and the Domuztepe excavations published through *OpenContext* which publishes research data on the web (**Gruber**, 2017). [Figure 22a and 22b]
https://opencontext.org

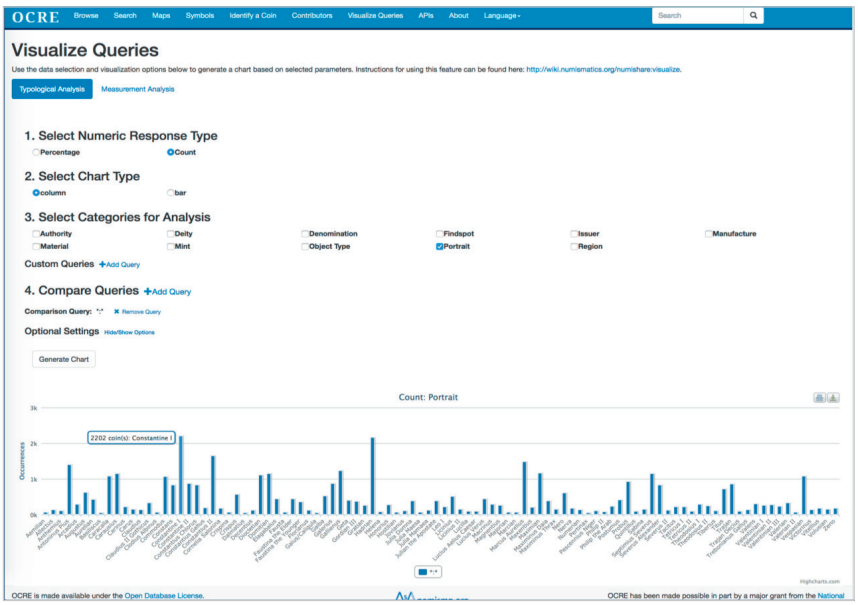


Figure 22c. *OCRE* visualize queries page.
http://numismatics.org/ocre/visualize

Even more useful to researchers is the visualized querying and analyzing across the whole dataset. Using the data selection and visualization options provided, a user can generate a chart based on selected parameters for typological analysis or measurement analysis. (Refer to the next figure.) [Figure 22c]

It is important to point out the meaning of this innovative resource to humanities researchers. End users are typically untrained in using LOD datasets through data dumps and Sparql endpoints (the most popular LOD deliverables). To master a query, one has to understand the syntax, forms, operators, result set modifiers, variables, and functions of the Sparql query language. Hence, easy access by end users becomes critical to the effective use of the LOD products. In order to better reuse digital resources while maximizing the outcomes for both machine and human users, easy-to-use LOD services have direct impacts on consuming these datasets as knowledge bases.

In the *OCRE* case, the visualize queries page (Figure 22c) allows the user to render queries in the form of charts and graphs, enabled by its JavaScript library at the back-end. Chart parameters are passed RESTfully by URL parameters, making it possible to bookmark and share charts over the web. Another powerful and useful analytic feature of *Numishare* is the maps interface. For example, the map page for hoard collections differs in one key way from other types of collections: The *Simile* timeline is incorporated into the interface through the TimeMap library. The points of the map correspond to find spots, and points in the timeline are created when hoard record contents contain datable coins.
http://wiki.numismatics.org/numishare:maps

Overall, users of *OCRE* are liberated from unfamiliar query languages and are provided effective analytics outcomes. This is the best example of how LAM data are used to support digital humanities researches. As **Gruber** (2017) pointed out, the study of coins has long been seen as an esoteric sub-discipline within history and archaeology, but the introduction of numismatics through intuitive user interfaces online serves as a bridge in exposing what is typically viewed as highly specialized information to a more general audience of archaeologists and classicists.

3.3.6. Discussion

It is obvious that there are limitless potentials for those [so far] un-structured data in LAMs to be used in supporting digital humanities research and education. This section showcased the pioneer products that have already brought LAM data to a totally new level of significance. It is also clear that Linked Data is a major concept that has been making semantic enrichment of a variety of LAM data in a distributed content creation environment possible and effective. Linked Data is about using the web to connect related data that wasn't previously linked, or using the web to lower the barriers to linking data currently linked using other methods.
http://linkeddata.org

The cases also imply that, semantic annotation, which formally identifies concepts and relations between concepts in documents, relies on both human and machine actions. Meanwhile, ontology-based approaches have already formed new mainstreams in digging into unstructured data and bringing unbelievable new discoveries to the front-end.

As pointed out by Robert Allen, some of the initial steps toward highly structured repositories may be relatively easy. However, scaling these will be difficult and requires active community engagement. Community involvement will also be required for policies and procedures in determining thresholds for consensus about results (**Allen**, 2017). It is obvious that the *IIIF* community introduced above is addressing such needs successfully and opportunely. Allen further suggested

that, for the representation of digital collections, rich semantic structures such as systems, models, simulations, events, and frames are needed (**Allen**, 2017). Each of the cases and research projects presented in this section so far have established their rich semantic structures at the back-end, delivering user-friendly, highly specialized information products to a more general audience at the front-end.

> "In data resources that are served through LAMs, unstructured data are usually available in the largest quantity in comparison with structured data, have the most diversity in type, nature, and quality, and are the most challenging to process. It is obvious that there are limitless potentials for those [so far] un-structured data in LAMs to be used in supporting digital humanities research and education"

Products such as *OCRE*, *Maphub*, and *Linked Jazz* can be considered as the representatives of Smart Data because they adequately represent a sufficient number of relevant features of humanistic objects of inquiry to enable the necessary level of precision and nuance required by humanities scholars, and also provide users with a sufficient amount of data to enable quantitative methods of inquiry, helping researchers to surpass the limitations inherent in methods based on close reading strategies (refer to **Schöch**, 2013). It is thrilling that LAM data providers and researchers in the humanities are incorporating the data-driven environment where advanced digital technologies have created the possibility of novel and hybrid methodologies.

3.4. Unifying heterogeneous contents in a distributed data creation environment

In digital humanities research, integration and interoperation of cultural heritage (CH) data are constantly in demand. An example was given by Prof. Eero Hyvönen in an invited talk at *VIII Encounter of documentation centres of contemporary art: open linked data and integral management of information in cultural centres*, held at the *Artium Museum*, Vitoria-Gasteiz, Spain, October 2016:

> "For example, if metadata about a painting created by Picasso comes from an art museum, it can be enriched (linked) with, e.g., biographies from *Wikipedia* and other sources, photos taken of Picasso, information about his wives, books in a library describing his works of art, related exhibitions open in museums, and so on. At the same time, the contents of any organization in the portal having Picasso related material get enriched by the metadata of the new artwork entered in the system. This is a win-win business model for everybody to join such a system; collaboration pays off. The linking can be established correctly only if unambiguous URI identifies are constantly used." (**Hyvönen**, 2016)

In order to achieve such a win-win result, it is important to realize the inherent needs and challenges. The heterogeneous contents could come with all kinds of formats and media, languages, cultural backgrounds, often companioned with provenance records and other contextual information. Yet another complicated situation is that their documentation has followed special professional standards and best practices defined and implemented in different professions involving libraries, archives, museums, and other cultural heritage institutions for a long time. The interoperability efforts of accommodating another community's model and data structure in one project might enable moving a step ahead, yet the uncertainty of quality and possible redundant tasks would be beyond prediction when diverse institutions are involved.

As a result, a fundamental semantic question in dealing with CH data is how to make the heterogeneous contents semantically interoperable, so that they can be searched, linked, and presented in a harmonized way across the boundaries of the datasets and data silos.

3.4.1 CASE: the *Sampo* portals

Hyvönen summarized three major semantic agreements that are needed for interoperability: (1) Domain neutral semantic model; (2) Metadata alignment model; and (3) Shared domain ontologies. The shared domain ontologies refer to the agreement of sharing domain ontologies (places, persons, etc.) whose concepts are used for populating the metadata models (**Hyvönen**, 2016).

In Figure 23, the data publication system is illustrated by a circle. A shared semantic ontology infrastructure is situated in the middle. It includes mutually aligned metadata and shared domain ontologies, modeled using Semantic Web standards. If content providers outside of the circle provide the system with metadata about a CH object, the data is automatically linked and enriched with each other and forms a *Giant Global Graph* (*GGG*) (**Hyvönen**, 2016).

The *Sampo* Model that has been tested and used in three cultural heritage case studies in Finland demonstrates such a win-win situation. *Sampo* is a mythical object in Finnish folklore that gives the holder wealth and good fortune. *https://en.wikipedia.org/wiki/Sampo*

In this instance, the *Sampo* model is a generic name for the model applied in various CH projects that created large-scale aggregated data sets for digital humanities applications from heterogeneous sources using Linked Data.

- *WarSampo*, the most notable and award-winning case,[19] was released in November 2015 and is the first large-scale system for serving and publishing WW2 LOD on the Semantic Web. The data draws eight different major datasets

from different organizations, originally totaling 7.6 million triples in a Sparql endpoint. The portal allows both historians and laymen to study war history and the destinies of their family members in the war from different interlinked applications such as Events, Persons, Army Units, Places, *Kansa Taisteli* magazine articles, Casualties, Photographs, and War Cemeteries. The data is annotated using a set of domain ontologies, including: 1) an ontology of the troops and their hierarchies, 2) persons with their ranks and roles, 3) place ontology of historical places, 4) event ontology of battles, politics, and other war time incidents, 5) an ontology of time periods, 6) ontology of weapons, 7) ontology of vessels, and 8) a subject matter ontology (**Hyvönen** *et al.*, 2016).
https://www.sotasampo.fi/en

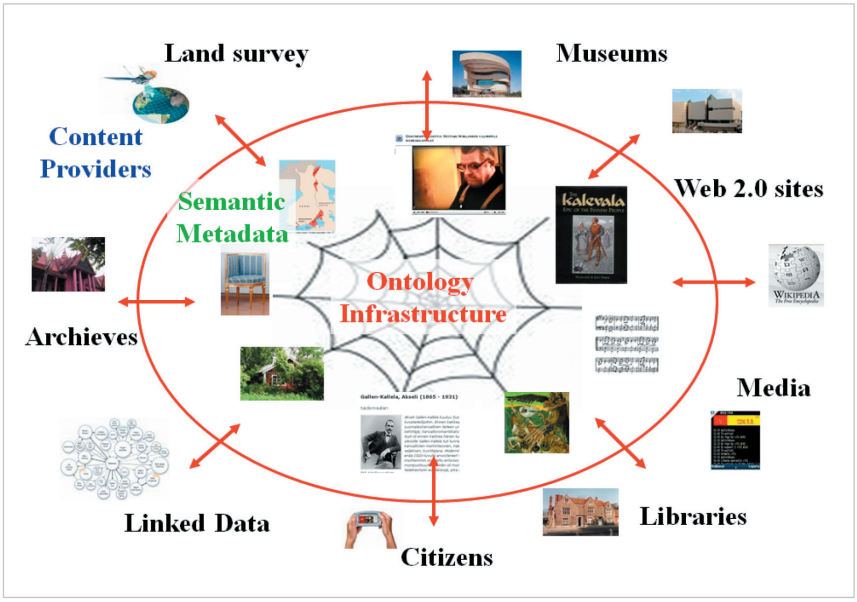


Figure 23. *Sampo* model for Linked Data publishing is based on a shared ontology infrastructure in the middle. Image source: **Hyvönen**, 2016

- *BiographySampo* is based on extracting knowledge from the underlying biographical texts —over 13,000 short biographies published by the *Finnish Literature Society*— using language technology, and by enriching the data through linking it to various external biographical databases, *Wikipedia*/*Wikidata*, collection databases of memory organizations, semantic web data services, etc. Similar to the *WarSampo* portal, users can find information via multiple different interlinked applications such as Persons, Places, Life maps, Statistics, Networks (of filtered people), Relations, and Language (**Hyvönen** *et al.*, 2018).
  http://biografiasampo.fi
- *NameSampo* for toponomastic research, is based on over two million places names collected in Finland and beyond. The objective of the project was to convert the place name entry slips, collection maps, and the attributes and metadata related to them into digital format (**Ikkala** *et al.*, 2018).
  https://seco.cs.aalto.fi/projects/nimisampo

The *Sampo* model has been applied to a series of semantic *Sampo* portals, resulting in more break-through LOD products. The portals include *CultureSampo* (2009), *BookSampo* (2011), *TravelSampo* (2011), *WarSampo* (2015), *NameSampo* (2018), and *BiographySampo* (2018).

3.4.2 Discussion

The *Sampo* model and the semantic portals demonstrated the win-win situation in dealing with CH data and making the heterogeneous contents of LAM data semantically interoperable, so that they can be searched, linked, and presented in a harmonized way across the boundaries of the datasets and data silos. They reveal a bright direction for the LAM data's next step in the semantic enrichment movement.

Parallel to the *Giant Global Graph* (*GGG*) featured by the *Sampo* model for Linked Data publishing —which is based on a shared ontology infrastructure—, integrating and reusing high level upper-ontologies and knowledge bases has led to another new product: *KBpedia*. Combining artificial intelligence (AI) with formal ontologies, the knowledge-based AI is pushing the speed of higher-level content analysis, entity recognition and categorization, and semantic network construction. A recently released open-source *KBpedia* built its knowledge structure by integrating seven core public knowledge bases — *Wikipedia*, *Wikidata*, *schema.org*, *DBpedia*, *GeoNames*, *OpenCyc*[20], and *Umbel*[21]. *KBpedia*'s upper structure, or knowledge graph, is the *KBpedia Knowledge Ontology* (*KKO*).
http://kbpedia.org

Written primarily in OWL 2, *KBpedia* includes 55,000 reference concepts, about 30 million entities, and 5,000 relations and properties, all organized according to about 70 modular typologies that can be readily substituted or expanded. Such a product characterizes the advanced semantic technologies that can be used for concept analysis and entity annotation, mapping, and data integration, plus automatic support for AI machine learning and semantic searching (*KBpedia*, 2018).

4. Summary and conclusion

With the rapid development of the digital humanities field, demands for smarter and bigger historical and cultural heri-

tage data, which usually cannot be obtained through web crawling or scraping, have focused attention toward LAM data, the treasure of all society. On the technology side, the semantic technologies and new artificial intelligence (AI) have brought innovative and implementable applications to the data-driven solutions. On the LAM side, the decade long digitization investment has formed LAM data as the most reliable resources for DH research (**Varner**; **Hswe**, 2016). Semantic enrichment is a strategy increasingly used during recent years, directly applied to the enhancement of LAM data (structured, semi-structured, and unstructured). In the "Discussion" sections dedicated to each of these data types in section 3, typical features and approaches are summarized and further discussions and resources are given. It is clear that there is no limit to the types of LAM data that can be semantically enriched. The real cases, research projects, experiments, and pilot studies shared in this article demonstrate endless potential for LAM data, whether they are structured, semi-structured, or unstructured, regardless of what types of original artifacts carry the data.

> Smart Data is a concept embraced by humanities research, and underlines the organizing and integrating processes from unstructured data to structured and semi-structured data, making the big data smarter (**Kobielus**, 2016; **Schöch**, 2013). Activities pertaining to LAM data enrichment have proved to be the initiator that enables LAMs to advance their data into smart data, supporting deeper and wider exploration and use of data in digital humanities research

Activities pertaining to LAM data enrichment have proved to be the initiator that enables LAMs to advance their data into smart data, supporting deeper and wider exploration and use of data in digital humanities research. The semantic enrichment strategy represents one major step and directly enhances LAM data by using semantic technologies. The cases and examples used in this article are representative of such activities. Following these roadmaps would encourage more effective initiatives and strengthen this effort to maximize LAM data's discoverability, use- and reuse-ability, and their value in the mainstream of DH and the Semantic Web.

Information technologies and new semantic technologies supporting each of the processes of semantic enrichment are advancing quickly. Reports of diverse experiments, tools, and issues can be found in publications across a wide range of domains, and in extraordinary depth. In addition to those discussed in the Discussion sections in section 3, AI-assisted processing is also becoming a norm for entity recognition, auto-recommendation, mapping, and verification. The fast extension of semantic technologies has been bringing astonishing news daily. Another closely related topic is increasing the findability of LAM data by exposing them to search engines, a benefit of semantic enrichment. Other articles in this journal will be dedicated to these topics, hence they are not covered in this article.

No matter what end products results from the semantic enrichment of LAM data, they will most likely join data on the web. Thus, this article concludes with the benchmarks recommended by the *W3C* in *Data on the web best practices*. The document focuses mainly on publishing data rather than consumption of data, and is geared toward data available through the web. It provides best practices related to the publication and usage of data on the web, so that data will be discoverable and understandable by humans and machines, while a self-sustaining ecosystem is facilitated. It is noteworthy that "provide metadata" is always recommended throughout the document. Each of the benchmarks also identify the ultimate goals for LAM data: comprehension, processability, discoverability, reuse possibility and effectiveness, trustiness, linkability, accessibility, and interoperability (**Farias-Lóscio**; **Burle**; **Calegari**, 2017). [Figure 24]

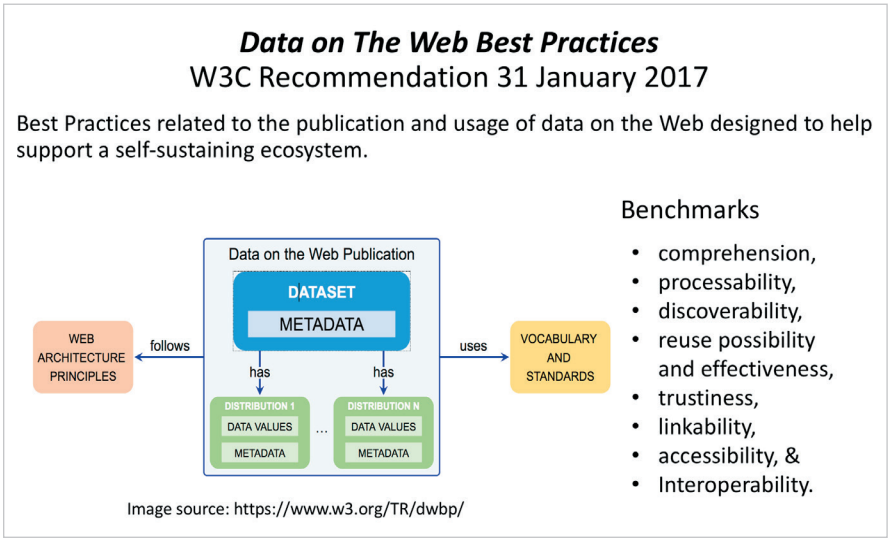


Figure 24.*W3C*'s data on the *Web best practices* recommendations and benchmarks.
Source: Created by the author based on the *W3C*, 2017.

Notes

1. As of 2016, participating nations and funding organizations include: Argentina (*MINCyT*); Brazil (*Fapesp*); Canada (*Sshrc*, *Nserc*, *FRQ*); Finland (*AKA*); France (*ANR*); Germany (*DFG*); Mexico (*Conacyt*); Netherlands (*NWO*); Portugal (*FCT*); United Kingdom (*AHRC*, *ESRC*), and United States (*NEH*, *NSF*, *IMLS*). Source: *https://diggingintodata.org/awards/2016/news/winners-round-four-t-ap-digging-data-challenge*

2. (*FRBR*) *Functional requirements for bibliographic records*

3. (*FRAD*) *Functional requirements for authority data*

4. (*FRSAD*) *Functional requirements for subject authority data*

5. *Europeana semantic enrichment* (November 5, 2015)
https://pro.europeana.eu/page/europeana-semantic-enrichment

6. Available from *Europeana* semantic enrichment website's link to "several vocabularies.
https://pro.europeana.eu/page/europeana-semantic-enrichment

7. According to April 19, 2018 note and updated Table 3 in **Manguinhas**, 2016.
http://shorturl.at/hyzL0

8. Data are available at:
https://www.kaggle.com/dorami/museum-project/data

9. *Getty Vocabularies: LOD* website
http://vocab.getty.edu

10. By clicking "View the full Getty record", a user is led to
http://vocab.getty.edu/page/ulan/500009365

11. *OpenRefine Reconciliation Service* (explanations and sample query templates).
http://vocab.getty.edu/queries#OpenRefine_Reconciliation_Service

12. *Museums and Collections with Maya Inscriptions*
http://mayawoerterbuch.de/museumscollections

13. Entry of *Museu Etnològic*
http://mayawoerterbuch.de/museums/museu-etnologic

14. Data collected on Oct. 23, 2018 from
https://old.datahub.io/dataset

15. "Data Sources. Oral History Transcripts."
https://linkedjazz.org/data-sources

16. Watch the video
http://maphub.github.io

17. About *IIIF*
https://iiif.io/about

18. Mogao Grottoes Cave 257, West Wall.
https://www.e-dunhuang.com/cave/10.0001/0001.0001.0257

19. *WarSampo* wins the *Open Data Prize* in the *2017 Lodlam Challenge*

20. *OpenCyc* (2002-2017) was a part of *Cyc* (/ˈsaɪk/), the world's longest-lived artificial intelligence project.
http://www.cyc.com

21. *Umbel* (*Upper Mapping and Binding Exchange Layer*)
http://umbel.org

5. References

Albritton, Benjamin (2013). *Digital manuscript interoperability: Shared canvas and IIIF in practice*.
https://slideplayer.com/slide/5840185

Alemu, Getaneh; **Brett, Stevens**; **Ross, Penny**; **Chandler, Jane** (2012). "Linked data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models". *New library world*, v. 113, n. 11/12, pp. 549-570.
http://eprints.rclis.org/17523
https://doi.org/10.1108/03074801211282920

Allen, Robert B. (2017). "Rich semantics and direct representation for digital collections". In: *ACM/IEEE Joint conference on digital libraries* (*JDCL*), pp. 348-349.
https://doi.org/10.1109/JCDL.2017.7991623

Appleby, Michael; **Crane, Tom**; **Sanderson, Robert**; **Stroop, Jon**; **Warnet, Simeon** (2012a). "IIIF Image API 2.1.1". IIIF.
https://iiif.io/api/image/2.1

Appleby, Michael; **Crane, Tom**; **Sanderson, Robert**; **Stroop, Jon**; **Warnet, Simeon** (2012b). "IIIF Presentation API 2.1.1". IIIF.
https://iiif.io/api/presentation/2.1

Bainbridge, David; **Hinze, Annika**; **Cunningham, Sally-Jo**; **Downie, J. Stephen** (2016). "Low-cost semantic enhancement to digital library metadata and indexing: Simple yet effective strategies". In: *2016 ACM/IEEE Joint conference on digital libraries* (*JDCL*), pp. 93-102.
https://core.ac.uk/download/pdf/44290466.pdf

Baker, Thomas; **Bermès, Emmanuelle**; **Coyle, Karen**; **Dunsire, Gordon**; **Isaac, Antoine**; **Murray, Peter**; **Panzer, Michael**; **Schneider, Jodi**; **Singer, Ross**; **Summers, Ed**; **Waites, William**; **Young, Jeff**; **Zeng, Marcia Lei** (2011). *Library linked data Incubator Group Final Report*. W3C Incubator Group Report 25.
http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025

Bensmann, Felix; **Zapilko, Benjamin**; **Mayr, Philipp** (2017). "Interlinking large-scale library data with authority records". *Frontiers in digital humanities*, n. 4, p. 5.
https://doi.org/10.3389/fdigh.2017.00005

Borgman, Christine L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press. ISBN: 978 0 262529914

Burdick, Anne; **Drucker, Johanna**; **Lunenfeld, Peter**; **Presner, Todd**; **Schnapp, Jeffrey** (2012). *Digital_Humanities*. Cambridge, MA: MIT Press. ISBN: 978 0 262528863

Clarke, David (2015). "Deep image annotation: Making a difference in knowledge organization". *Fourth ISKO-UK Biennial conference of the International Society for Knowledge Organization*.
http://docplayer.net/13812285-Deep-image-annotation-making-a-difference-in-knowledge-organization.html

Consultative Committee for Space Data Systems (2012). *Reference model for an Open Archival Information System*. Washington DC: CCSDS.
https://public.ccsds.org/Pubs/650x0m2.pdf

Damjanovic, Violeta; **Kurz, Thomas**; **Westenthaler, Rupert**; **Behrendt, Wernher**; **Gruber, Andreas**; **Schaffert, Sebastian** (2011). "Semantic enhancement: The key to massive and heterogeneous data pools". In: *Proceedings of the 20th intl IEEE ERK* (*Electrotechnical and Computer Science*) conference, pp. 413-416.
https://www.researchgate.net/publication/266603290_Semantic_Enhancement_The_Key_to_Massive_and_Heterogeneous_Data_Pools

Dunsire, Gordon; **Willer, Mirna** (2011). "Standard library metadata models and structures for the semantic web". *Library hi tech news*, v. 28, n. 3, pp. 1-12.
https://doi.org/10.1108/07419051111145118

Farias-Lóscio, Bernadette; **Burle, Caroline**; **Calegari, Newton** (2017). *Data on the web best practices. W3C Recommendation 31 January 2017*.
http://www.w3.org/TR/dwbp

Floridi, Luciano (2010). *Information: A very short introduction*. Oxford: Oxford University Press. ISBN: 978 0 199551378

Gardner, Dan (2012). "An ocean of data [Introduction]". In: Smolan, Rick; Erwitt, Jennifer (eds.). *The human face of big data*. Sausalito, CA: Against All Odds Productions, pp. 14-17. ISBN: 978 1 454908272

Gracy, Karen; **Davidson, Sammy** (2014). "Helping users find the 'good stuff': Using the semantic analysis method (SAM) tool to identify and extract potential access points from archival finding aids". In: *SAA Research Forum*, Society of American Archivists.
http://files.archivists.org/pubs/proceedings/ResearchForum/2014/posters/GracyDavidson-ResearchForumPoster2014.pdf

Gracy, Karen; **Zeng, Marcia Lei** (2015). "Creating linked data within archival description: Tools for extracting, validating, and encoding access points for finding aids". *Digital humanities conference of the Alliance of Digital Humanities Organizations* (*ADHO*).

Gracy, Karen; **Zeng, Marcia Lei**; **Skirvin, Laurence** (2013). "Exploring methods to improve access to music resources by aligning library data with linked data: A report of methodologies and preliminary findings". *Journal of the American Society for Information Science and Technology* (*JASIS&T*), v. 64, n. 10, pp. 2078-2099.
https://doi.org/10.1002/asi.22914

Gruber, Ethan (2017). "Final report to the NEH for online coins of the Roman Empire". *Day of archaeology*, July 28.
http://www.dayofarchaeology.com/final-report-to-the-neh-for-online-coins-of-the-roman-empire

Hinze, Annika; **Taube-Schock, Craig**; **Bainbridge, David**; **Matamua, Rangi**; **Downie, J. Stephen** (2015). "Improving access to large-scale digital libraries through semantic-enhanced search and disambiguation". In: *Proceedings of the 15th ACM/IEEE-CS Joint conference on digital libraries*. Association for Computational Linguistics, pp. 147-156.
https://doi.org/10.1145/2756406.2756920

Hyvönen, Eero (2016). "Cultural heritage linked data on the semantic web: Three case studies using the sampo model". *VIII Encounter of documentation centres of contemporary art: open linked data and integral management of information in cultural centres*. Artium, Vitoria-Gasteiz, Spain, October 19-20.
https://seco.cs.aalto.fi/publications/submitted/hyvonen-vitoria-2017.pdf

Hyvönen, Eero; **Heino, Erkki**; **Leskinen, Petri**; **Ikkala, Esko**; **Koho, Mikko**; **Tamper, Minna**; **Tuominen, Jouni**; **Mäkelä, Eetu** (2016). "Publishing Second World War history as linked data events on the semantic web". In: *Proceedings of the digital humanities conference*, pp. 571-573.
https://seco.cs.aalto.fi/publications/2016/hyvonen-et-al-warsa-dh2016.pdf

Hyvönen, Eero; **Leskinen, Petri**; **Tamper, Minna**; **Rantala, Heikki**; **Tuominen, Jouni**; **Keravuori, Kirsi** (2018). "Demonstrating BiographySampo in solving digital humanities research problems in biography and prosopography" [Submitted paper].
https://seco.cs.aalto.fi/publications/submitted/hyvonen-et-al-bs-2019.pdf

Ikkala, Esko; **Tuominen, Jouni**; **Raunamaa, Jaakko**; **Aalto, Tiina**; **Ainiala, Terhi**; **Uusitalo, Helinä**; **Hyvönen, Eero** (2018). "NameSampo: A linked open data infrastructure and workbench for toponomastic research". In: *GeoHumanities 18*, *Proceedings of the 2nd ACM SIG Spatial workshop on geospatial humanities*, Seattle, WA, USA, November 06-09, pp. 2:1-2:9, ACM.
https://doi.org/10.1145/3282933.3282936

IMLS (2018). *Transforming communities: IMLS strategic plan (2018-2022)*. Washington DC: Institute of Museum and Library Services.
https://www.imls.gov/sites/default/files/publications/documents/imls-strategic-plan-2018-2022.pdf

Isaac, Antoine; **Manguinhas, Hugo**; **Stiller, Juliane**; **Charles, Valentine** (2015). *Report on enrichment and evaluation*. The Hague, Netherlands: Europeana Task Force on Enrichment and Evaluation.
http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf

Kaplan, Frederic (2015). "A map for big data research in digital humanities". *Frontiers in digital humanities*, n. 2.
https://doi.org/10.3389/fdigh.2015.00001

KBpedia (2018). *KBpedia is now open source*, October 23.
http://kbpedia.org/resources/news/kbpedia-is-open-source

Kobielus, James (2016). "The evolution of big data to smart data". In: *Smart data online*, July 13 [PowerPoint slides].
http://smartdata2016.dataversity.net

Lin, Yuri; **Michel, Jean-Baptiste**; **Lieberman-Aiden, Erez**; **Orwant, Jon**; **Brockman, Will**; **Petrov, Slav** (2012). "Syntactic annotations for the Google Books Ngram corpus". In: *Proceedings of the ACL 2012 System demonstrations*. Association for Computational Linguistics, pp. 169-174.
http://aclweb.org/anthology/P12-3029

Manguinhas, Hugo (ed.) (2016). *Europeana semantic enrichment framework*. Version 17, Nov.
https://pro.europeana.eu/page/europeana-semantic-enrichment
http://shorturl.at/pEIQ5

Mayer, Daniel (2011). *Mainstream semantic enrichment* [*YouTube* video]. December 26.
http://www.youtube.com/watch?v=YVxvQ7UpqI0

Mayer-Schönberger, Viktor; **Cukier, Kenneth** (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Eamon Dolan/Mariner Books. ISBN: 978 0 544227750

MoMA (2017). *Museum of Modern Art Collection -- Title, artist, date, and medium of every artwork in the MoMA collection*, Last Updated: 2017 (Version 1).
https://www.kaggle.com/dorami/museum-project/data

Mukerjee, Prithwis (2014). "Introduction to data science" [*PowerPoint* slides], January 12.
http://www.slideshare.net/prithwis/01-intro2-datascienceyantrajaalblog

Mutuvi, Stephen; **Doucet, Antoine**; **Odeo, Moses**; **Jatowt, Adam** (2018). "Evaluating the impact of OCR errors on topic modeling". In: *Maturity and innovation in digital libraries. 20th Intl conf on Asia-Pacific digital libraries*, *ICADL 2018*, Ha-

milton, New Zealand, November 19-22, Proceedings, pp. 3-14. ISBN: 978 3 030 04257 8

National Archives (2016). "Finding aid type". *The national archives catalog*.
https://www.archives.gov/research/catalog/lcdrg/elements/findingtype.html

Nguyen, Thi-Tuyet-Hai; **Coustaty, Mickael**; **Doucet, Antoine**; **Jatowt, Adam**; **Nguyen, Nhu-Van** (2018). "Adaptive edit-distance and regression approach for post-OCR text correction". In: *Maturity and innovation in digital libraries. 20th Intl conf on Asia-Pacific digital libraries*, *ICADL 2018*, Hamilton, New Zealand, November 19-22, Proceedings, pp. 278-289. ISBN: 978 3 030 04257 8

O'Neill, Ed; **Mixter, Jeff** (2013). "Maximizing the usage of value vocabularies in the linked data ecosystem". In: *76th Annual meeting of the American Society for Information Science and Technology* (*ASIS&T*), Montreal, Canada, November.
http://nkos.slis.kent.edu/ASIST2013/ONeill-Mixter.pptx

Pattuelli, M. Cristina (2012). "Personal name vocabularies as linked open data: A case study of jazz artist names". *Journal of information science*, v. 38, n. 6, pp. 558-565.
https://doi.org/10.1177/0165551512455989

Pattuelli, M. Cristina; **Hwang, Karen**; **Miller, Matthew** (2016). "Accidental discovery, intentional inquiry: Leveraging linked data to uncover the women of jazz". *Digital scholarship in the humanities*, v. 32, n. 4, pp. 918-924.
https://doi.org/10.1093/llc/fqw0

Prasad, A. R. D.; **Giunchiglia, Fausto**; **Devika, P. Madalli** (2017). "DERA: from document centric to entity centric knowledge modelling". In: *Proceedings of the International UDC seminar 2017. Faceted classification today*. London, September, pp. 169-179.
http://seminar.udcc.org

Riva, Pat; **LeBoeuf, Patrick**; **Žumer, Maja** (2017). *IFLA library reference model: A conceptual model for bibliographic information*. Netherlands: IFLA.
https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf

Schöch, Christof (2013). "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of digital humanities*, v. 2, n. 3, pp. 2-13.
http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities

Smith-Yoshimura, Karen (2018). "The rise of Wikidata as a linked data source". In: *Hanging together. The OCLC research blog*, August 6.
http://hangingtogether.org/?p=6775

Stiller, Juliane; **Petras, Vivien**; **Gäde, Maria**; **Isaac, Antoine** (2014). "Automatic enrichments with controlled vocabularies in Europeana: Challenges and consequences." In: *Euro-Mediterranean conf.*, pp. 238-247. Springer, Cham.
https://doi.org/10.1007/978-3-319-13695-0_23

Svensson, Patrik (2010). "The landscape of digital humanities". *Digital humanities quarterly*, v. 4, n. 1.
http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html

Thorsen, Hilary K.; **Pattuelli, M. Cristina** (2016). "Linked open data and the cultural heritage landscape". In: Jones, Ed; Seikel, Michele (eds.). *Linked data for cultural heritage*. Chicago, IL: Alcts Publishing. ISBN: 978 1 783301621

TiECON East (2014). *Data is the new oil*.
https://tieconeast.wordpress.com/page/2

Reinhard, Andrew; **Van-Alfen, Peter**; **Bransbourg, Gilles**; **Gruber, Ethan**; (2017). "Wishes granted: the ANS and the NEH". In: *National Endowment for the Humanities. Announces. New grant recipients*.
http://numismatics.org/pocketchange/wp-content/uploads/sites/3/NEH-Article-ANS-Magazine.pdf

Van-Ruyskensvelde, Sarah (2014). "Towards a history of e-ducation? Exploring the possibilities of digital humanities for the history of education". *Paedagogica historica*, v. 50, n. 6, pp. 861-870.
https://doi.org/10.1080/00309230.2014.955511

Varner, Stewart; **Hswe, Patricia** (2016). *Special report: Digital humanities in libraries*. American Libraries.
https://americanlibrariesmagazine.org/2016/01/04/special-report-digital-humanities-libraries

W3C (2011). Library Linked Data Incubator Group Final Report
https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025

W3C (2017). Data on the Web best practices.
https://www.w3.org/TR/dwbp

Wagner, Elisabeth; **Matsumoto, Mallory**; **Kiel, Nikolai**; **Gronemeyer, Sven** (2014). *A checklist of museums with Maya Art*.
http://mayawoerterbuch.de/museumscollections

Wang, Xiaoguang; **Liu, Xuemei**; **Xia, Shengping** (2017). "Design and implementation of deep semantic indexing on digital cultural heritage images". *Journal of library and information science*, v. 43, n. 1, pp. 98-121.
http://jlis.glis.ntnu.edu.tw/ojs/index.php/jlis/article/view/716

Weitz, Jay; **Toves, Jenny**; **Vizine-Goetz, Diane**; **Naught, Nannette**; **Bremer, Robert** (2016). "Mining MARC's hidden treasures: Initial investigations into how notes of the past might shape our future". Journal of library metadata, v. 16, n. 3-4, pp. 166-180.
https://doi.org/10.1080/19386389.2016.1262653

Zeng, Marcia Lei (2017). "Smart data for digital humanities". *Journal of data and information science*, v. 2, n. 1, pp. 1-12.
https://doi.org/10.1515/jdis-2017-0001

Zeng, Marcia Lei; **Gracy, Karen**; **Skirvin, Laurence** (2013). "Navigating the intersection of library bibliographic data and linked music information sources: A study in the identification of useful metadata elements for interlinking". *Journal of library metadata*, v. 13, n. 2-3, pp. 254-278.
https://doi.org/10.1080/19386389.2013.827513

Zeng, Marcia Lei; **Gracy, Karen F.**; **Žumer, Maja** (2014). "Using a semantic analysis tool to generate subject access points: A study using Panofsky's theory and two research samples". *Knowledge organization*, v. 41, n. 6, pp. 440-451.
https://pdfs.semanticscholar.org/bbeb/42b931fd32520a03167770d2b5de694128e6.pdf

Zeng, Marcia Lei; **Mayr, Philipp** (2018). "Knowledge organization systems (KOS) in the semantic web". *International journal on digital libraries*.
https://rdcu.be/PgZW
https://doi.org/10.1007/s00799-018-0241-2

Žumer, Maja (2018). "IFLA library reference model (IFLA LRM): Harmonisation of the FRBR family". *Knowledge organization*, v. 45, n. 4, pp. 310-318.

Also available in **Hjørland, Birger** (ed.). *ISKO Encyclopedia of knowledge organization*.
http://www.isko.org/cyclo/lrm

Žumer, Maja; **Riva, Pat** (2017). "IFLA LRM-Finally here". In: *Intl conf on Dublin Core and metadata applications*, Washington, D.C., USA, 26-29 October, pp. 13-23.
http://dcpapers.dublincore.org/pubs/article/download/3852/2037