Moving EMLoT towards the web of data: an approach to the representation of humanities citations based on role theory and formal ontology

By Michele Pasin and Riichiro Mizoguchi
Department of Digital Humanities, King's College London
Mizoguchi Lab, Osaka University, Japan

1 Introduction: towards an interconnected web of data

The World Wide Web in the last two decades has undergone a great revolution due to the massive amount of both novel and pre-existing materials that have been digitized and made available in the form of online databases and, more generally, websites. It is now often the case that, despite this wealth of information is indeed available somewhere, it cannot be easily put to work so to make new and interesting research and discoveries, because it is unconnected, fragmented, or incompatible in the forms it has been digitized into. This leads to a scenario in which information repositories are resemblant of islands that are 'unaware' of each other, for they cannot be connected without dedicated manual and time-consuming work. Traditional search technologies based on string similarity and statistical methods (e.g., Google) have gone a long way in giving us a sense for how vast the web is; nonetheless, they reveal their limitations when it comes to merging data and connecting them together in a way that adds value to the overall result.

In order to address this problem, enterprises such as the Semantic Web (Berners-Lee et al. 2001) and more recently, the Linked Data initiative (Heath & Bizer 2011), have worked towards the construction of technical solutions that make it easier for users to contribute to and benefit from this expanding new ecosystem of online data - the *web of data* - as if it were just an extension of the usual web we all are already so accustomed to. Examples of the use of these solutions can be found in numerous areas. On the academic side, for example, projects such as the CIDOC-CRM (Doerr 2003) have been active in the task of creating a network of interlinked data repositories for the cultural heritage sector. Less academic examples are instead Freebase[1], a project aiming at becoming an "open, shared database of the world's knowledge" which can be freely edited by registered users, and the DBpedia (Auer et al. 2007), a community effort to extract structured information from Wikipedia and make it available as if it were a giant and constantly updated web database.

[1] www.freebase.com

In this paper, we refer to these recent developments in web technologies with the term ‘semantic web’ (SW), as they all share the intent to formally encode aspects of the *meaning* of the resources or artifacts they intend to describe (with varying degree of complexity and precision). Arguably, the key idea behind this approach is that by making available both the data and (one or more) formal descriptions of their meaning in the form of a suitable knowledge representation language (Davis et al. 1993), new and exciting things will soon or later happen thanks to a combination of enabling factors that include standard data integration technologies, advanced artificial intelligence techniques, and, last but not least, the ‘snowball’ effect deriving from the mere scale of the web.

It is important to underline that what the semantic web aims at achieving goes well beyond the mere *syntactic* integration of structured data (Uschold 2003). We normally talk about syntactic interoperability, for example, when two sources both make available their data using the same format, e.g., XML or a relational database language. Although this level of interoperability is necessary, it is not enough for the semantic web to work correctly, since these data representation technologies still require some level of human intervention in order to be integrated correctly. That is to say, programmers (and more generally, computer users) still play a very important role, which is similar to that one of an ‘interpreter’ that by drawing on various sources, understands the implicit and explicit semantics of two data formats and thus proceeds to the creation of a workable mapping between them. The results of this act of interpretation can then be embedded in the application layer of a database-backed website, or in the XSLT stylesheet that transforms an XML file into the final document we can read online. In other words, the meaning of the data (that is, what lets us manipulate and merge them effectively) is either not explicitly encoded, or it is encoded using a non-standardized knowledge representation language: in both cases, some key pieces of information remain implicit and thus they are essentially invisible to the machine (Cover 1998).

For example, we can imagine a situation in which two organizations that publish data about theatrical troupes decide to encode it in XML but ends up using different XML tags (e.g., <troupe> and <company>): if the intended meaning of these tags has been described accurately enough in the accompanying documentation, people could easily infer that in both cases we are talking about troupes in the same sense, and thus safely bring together the two datasets to the purpose of, say, a federated search application. Obviously, this is not achievable by computers alone, since computers have no means of understanding surrogate information like documentation and source-code comments. As a result, the two data sources about troupes could not be integrated automatically unless the intended meaning of their encoding is disambiguated by us. If we think of the scale of the web, it is not difficult to realize that this is one of the most pressing

problems that divide us from a web of isolated data-islands, to a web of interconnected information.

The semantic web vision proposes a number of solutions to at least ameliorate this problem (for a more detailed introduction, see (Palmer 2001)). Among them we have technologies such as Uniform Resource Identifiers (URIs) for creating stable and *dereferenceable*[2] identifiers of the objects described in our data, and formal *ontology* languages such as RDF and OWL in order to encode the semantics of our data (i.e., the assertions and background knowledge implicit in our data). In particular, this second point is not just about using a new fancy computer language, but it is primarily about following the principles of ontology engineering (Gruber 1995) when modeling the meaning of our data (principles that, it is worth noting, derive from the body of work of philosophical ontology itself (Poli & Obrst 2009)). Only by doing so, we can build the infrastructure needed for moving from a simple syntactic integration, to a semantic one, that is, one based on a more transparent and application-independent representation of the meaning of the resources being shared.

So, for example, if our two data providers deal with theatrical troupes, by following the semantic web approach they would attempt to publish their data using a common RDFS/OWL ontology, or at least, one that contains common elements or can be (at least partially) mapped to other ontologies available on the semantic web.

It is worth noting that this is, to some degree, very similar to what mechanisms such as standard *metadata* (Garshol 2004)set out to achieve. And indeed, there definitely is a commonality of purposes, in so far as both approaches attempt to facilitate finding and connecting information. The main difference with traditional metadata, it is worth pointing out, resides in the fact that the interoperability of metadata stops at the syntactical level. In other words, traditional metadata still need to be interpreted by humans in order to be used effectively in a data integration scenario: semantic technologies, on the other hand, encode the meaning of data using languages that have a clear formal semantics, thus allowing a much wider range of automatic and semi-automatic computing applications to be built on top of them.

1.1 Towards a web of interconnected bibliographical data

The application of ontology-based technologies to the bibliography discipline (intended as the academic study of books as physical and cultural objects) has long been the object of interest for researchers in the semantic web area.

Approaches such at BIBO (D'Arcus & Giasson 2009) or CITO (Shotton 2009) aim at providing the basic formal conceptual framework for distinguishing the main elements necessary

[2] In the web of data, a URI is dereferenceable if we can obtain a copy or representation of the resource it identifies using any of the internet protocols (e.g. HTTP).

to describe a document (such as *title*, *author, reference* or *publication-type*) or the relations existing between different documents (for example, *contradicts* or *uses-as-example*). Making explicit this type of knowledge by using semantic technologies allows to perform complex query lookups across different repositories, integrate data sources, and build visualizations or statistical analyses that expose the knowledge networks that exist in the literature. The two ontologies just mentioned are not the only examples of such approaches; many others are available that vary in the degree of precision they possess in modeling bibliographical data: (Portal 2002), (Gruber 1992), (Kruk et al. 2005).

However, despite the proliferation of formal models for capturing bibliographical information there are still aspects of this domain that need ontological clarification. In particular, we argue that this is due to the fact that most of existing research on this topic originates from scientific contexts, or, at best, ones that are not humanistic. This conclusion can be drawn - for example - from an ontological analysis of the common methods of analyzing and reconstructing the network of relationships among texts in the humanities.

The study of how texts quote and reference each other is a fundamental aspect of humanities disciplines at large; more specifically, it is the main goal of *textual criticism*. Scholars in this field of study aim at the identification of the elements that let them distinguish an original document (e.g., a manuscript) from its subsequent copies, to the purpose of better understanding phenomena such as the history of transmission of a text, or the propagation of an idea across the time and space dimensions. Accordingly, structural elements such as *citations* and *quotations* become objects of study in themselves, for they can reveal much more than a simple conceptual reference to a preexisting document.

In this essay we present an approach based on role theory and formal ontology whose purpose is to represent the key concepts in the field of textual transmission history. By characterizing the semantics of concepts such as *reference*, *citation*, *quotation* as they are normally used in this academic discipline, we aim at providing a framework that can enable more precise information integration services in this area, and more generally, in the digital humanities. At the same time, we want to show that the formalization of citations objects in the humanities is more complicated that expected, and that it requires an extension (and possibly a revision) of its correspondent solutions in the scientific world.

Our research derive from the *Early Modern London Theatres* project (EMLoT), whose goal was to build a database that lets you see what direct use has been made, over the last four centuries, of pre-1642 documents related to professional performance in purpose-built theaters and other permanent structures in the London area. In what follows we present our attempt to express the meaning of EMLoT's data using a formal ontology, we discuss the methods being used and, more generally, highlight the various difficulties involved in the formal modeling of the

domain of textual transmission history. The ontology will serve two purposes: first, it clarifies the meaning of the relevant concepts in this area; second, it poses the basis for further research aimed at facilitating the integration of this dataset with other ones available online.

The rest of the article is thus subdivided: section 2 gives more information about the project our research questions originated from; in section 3 and 4 we outline the desiderata of this work, and summarize generic methodology used to achieve them; finally, section 5 and 6 present the results we achieved and discusses them in the light of future research.

2 Background: the EMLoT project

The Early Modern London Theatres project[3] is a website and database resulting from an international collaboration between Toronto's Records of Early English Drama research centre and London's Department of Digital Humanities. The stated purpose of EMLoT is to identify, record and assess transcriptions from primary-source materials relating to the early London stage, as found in secondary-source print and manuscript documents (MacLean et al. 2012). The main criterion in distinguishing between a primary- and secondary-source document is chronological: the EMLoT's purview stops at 1642 (a critical year for theatre historians, for in that year London's theaters were closed by the rising Puritan movement, under accusations of moral depravation). Therefore, a primary source is a document produced before 1642, and a secondary source is one produced after 1642. Furthermore, in parallel with EMLoT's intended goal of tracing the history of transmission of primary sources, we also recorded information about the contents of such documents, that is, their reference to events of other factual information relevant to the theatrical scene or context at large, such as people (players, patrons), troupes and venues.

Fig. 1 contains a schematization of the main entities represented in the EMLoT database. These can be grouped into the *textual* pole, containing standardized bibliographical information about the primary and the secondary source, and the *factual* pole, containing information about the events described by the documents in questions, and related entities. Finally, on a plane orthogonal to these entities lies the 'Transcription Record' pole, which joins together a primary and secondary source document, and provides data concerning that relationship: the location of the transcription within the secondary document; citation data for the primary document (as

3 www.emlot.kcl.ac.uk

provided by the secondary source); brief notes on the treatment of the primary by the secondary source.

The 'Transcription Record' entity can be seen as the reification of the bibliographer's activity - the eye of the beholder - who identifies a transcription and provides some interpretive data about it. This information structure is extremely important, since it reflects a key interest of the researchers who composed the EMLoT database. In fact, with EMLoT we sought to provide as complete a profile as possible, not only of the documentary history of the early London stage, but also of the treatment of each document comprising this history. Thus we may encounter transcriptions that are faithful in every respect to the original, but also ones that are excerpted, emended and otherwise adulterated.

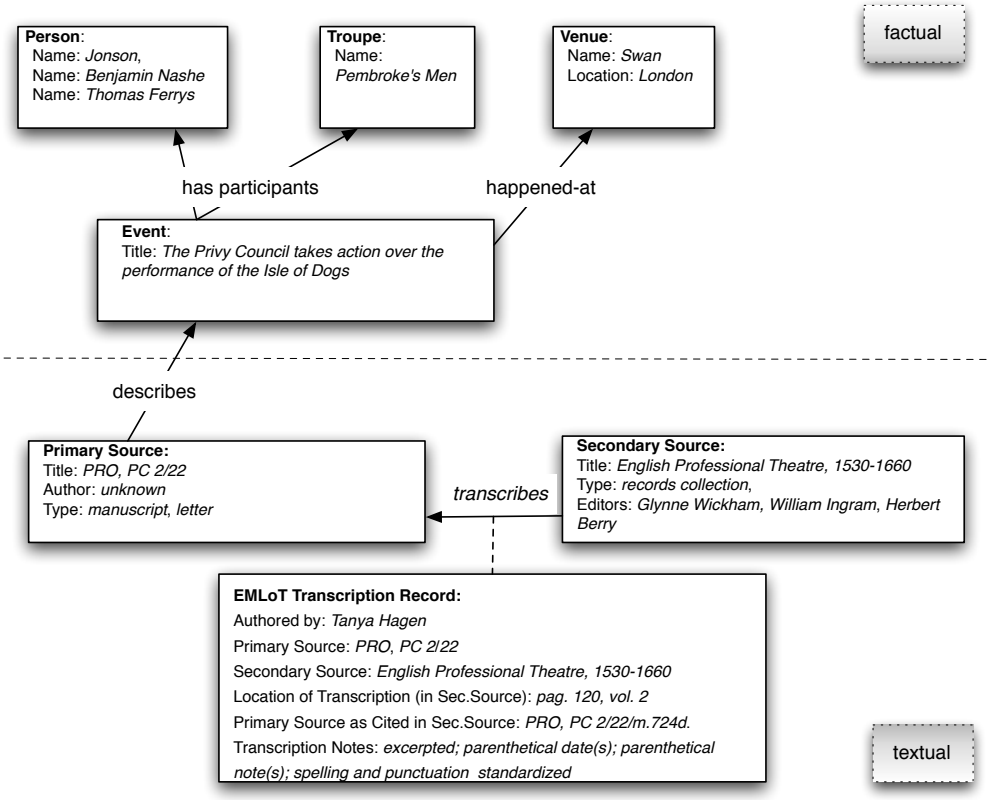

Fig. 1. Schematic representation of the information objects contained in the EMLoT database

All of these cases are equally important from the point of view of textual transmission history, as they will eventually let us consider not only the frequency with which a primary-source document has been published, but also its various treatments over time, and at the hands of different editors. Which documents tend to be preserved whole, and which heavily excerpted; which preserved in facsimile, and which modernized? The signal purpose of EMLoT, then, is to identify a relationship between a primary (transcribed) and a secondary (transcribing) document,

and the job of compiling the EMLoT database is to describe that relationship within the parameters of an established template.

It is in the *triangulation* of record and document files that the EMLoT serves both its nominal function as a bibliographic resource, and an ancillary historiographical purpose. And it is precisely this level of abstraction what we aimed at clarifying through the subsequent ontological analysis.

3 Constructing an ontology for EMLoT: research objectives

In this section we outline the main desiderata of this research work. In general, our main purpose is to design a formal ontological model that reflects the semantics of the database contents in a more transparent and generic fashion. Thanks to this ontology, we will be able to facilitate programmatic access and reuse of the EMLoT data; furthermore, we also would like to stimulate more debate and collaboration among researchers involved in the construction of similar digital resources in the humanities.

With reference with the distinction made in the previous section between the *factual* and the *textual* poles, we immediately realized that the information objects within the factual pole are quite commonly found in other digital projects too, and that adequate ontological schemas already exist for them. In such cases we decided to reuse what is already available, and avoid creating yet another overlapping conceptual model. In particular, we applied this approach in the following situations:

1) With entities like *persons* (“Benjamin Nashe”), *groups of people* or *organizations* (“Penbroke’s Men”) and *theaters/institutions* (“Blackfriars”). These can be easily represented using the FOAF vocabulary (Brickley & Miller 2010), or the CIDOC-CRM ontology (Crofts et al. 2008). Also, we found classes useful for representing theatrical groups on schema.org (`http://schema.org/TheaterGroup`) and the theatre ontology (`http://purl.org/theatre`).
2) With entities describing *place-names* (“the bear garden”) and their geographical *locations* (“Southwark”). A number of ontologies include concepts for represented places. For our purposes, we found the conceptualization offered in CIDOC-CRM to be good enough (cf. the `http://purl.org/NET/cidoc-crm/core#E53_Location` class, as described in (Crofts et al. 2008)on pp. *xiii* and 21).
3) With entities describing historical *events*. EMLoT includes a quite sophisticated categorization of events, in the form of an ‘event-type’ authority list. However, this list

is used more like a source of 'tags' for better searching events objects, rather than as a way to characterize events at the structural level. As a result, it is not easy to create a valid taxonomy from the event-types authority list, without the risk of misrepresenting its original purpose. In order to address this situation, we decided to formalize all event objects using the CIDOC-CRM event class (`http://purl.org/NET/cidoc-crm/core#E5_Event`), then describe them further using the data in the event-type authority list by means of the 'has-type' (`http://purl.org/NET/cidoc-crm/core#P2_has_type`) construct. This is a mechanism in CIDOC-CRM thatt allows the association of arbitrary free-form descriptors to any entity in a model (cf. (Crofts et al. 2008) pp. xi-xii).

The task of re-expressing EMLoT's data using a more principled ontological approach became most interesting as soon as we attempted to model the key structural elements of textual transmission history, as they emerge from the way data had been collected and organized in the database (cf. also fig. 2).

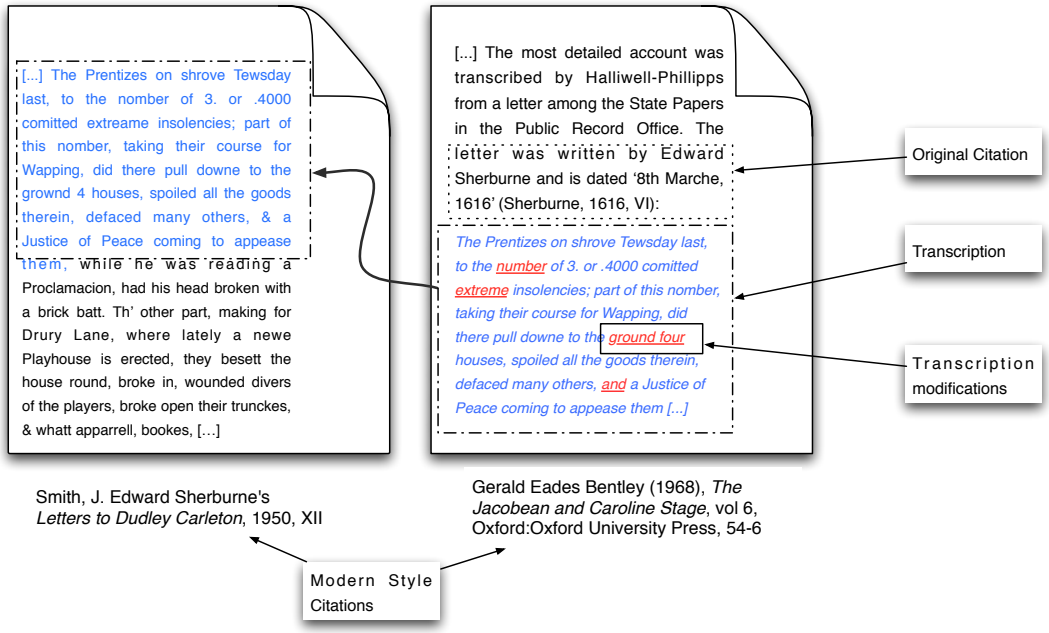

Fig 2. Representation of key structural elements in textual transmission history.

For example, unusual modeling challenges arise from the fact that researchers are not only interested in knowing that some (secondary) documents cite some other (primary) documents, but are equally interested in:

a) how the citation object is structured and formatted;

b) whether we are in the presence of just a citation, of a transcription, or a combination of the two;

c) whether the citation and/or transcription is genuine;

d) creating a new record, which is itself citable, containing the results of their analysis.

The remarkable character of these 'citations within citations' (a phenomenon which is not at all uncommon in the humanities) becomes apparent in the transcription record, the database object where these connections are stored (as discussed in section 2). A transcription record contains modern citation information about the primary and secondary source; information about the text being transcribed; finally, it also contains information about the citation style used in the secondary source when citing or quoting the primary source, and about errors or interesting changes of the transcription.

All of these different features are important because they may shed light on aspects of the history of the transmission of these documents. The ontological model should therefore provide a language that can describe them at the required granularity.

What emerges from this scenario is that we have different type of citations objects, each one with its own distinctive structure. To make things even more difficult, many of the key terms in textual transmission history are often not used in the same sense (cf. Table 1).

Term	Explanation
Reference	*noun* **1** the action of mentioning or alluding to something*: he **made reference to** the enormous power of the mass media | references to Darwinism and evolution*. • a mention or citation of a source of information in a book or article. • a book or passage cited in such a way.
Citation	*noun* **1** a quotation from or reference to a book, paper, or author, esp. in a scholarly work*: there were dozens of citations from the works of Byron | recognition through citation is one of the principal rewards in science*. • a mention of a praiseworthy act or achievement in an official report, esp. that of a member of the armed forces in wartime.
Transcription	*noun* a written or printed representation of something. • the action or process of transcribing something: *the funding covers transcription of nearly illegible photocopies*.
Quotation	*noun* **1** a group of words taken from a text or speech and repeated by someone other than the original author or speaker *: a quotation from Mark Twain | biblical quotations*. • the action of quoting from a text, speech, piece of music, or work of art *: a great argument with much quotation of Darwin*.

Table 1. Definition of key terms in textual transmission history (adapted from the New Oxford American Dictionary)

Rather, their intended meaning varies depending on the context, or on the particular background of the speaker. We are referring here to very commonly used words such as *reference*, *citation*, *transcription* and *quotation*. Since in everyday language the meanings of such terms tend overlap with each other, the task of creating a formal model for textual transmission history will also entail providing an unambiguous definition of these terms[4].

In conclusion, to the end of bringing some conceptual order to what seems, at first sight, an inexplicable and recursive state of things, in the following analysis we aim at:

a) outlining an ontological approach that can be used as a reference framework for the formalization and sharing of data in textual transmission history, and
b) showing how the various textual entities arising in EMLoT can be reduced to and represented through this logical model.

4 Methods

In this section we are going to spend a few words on the specific methods employed to carry out our ontological analysis, for they had a considerable effect on the results we achieved. In particular, our work relied on two choices. The first one, more general, concerns the decision of using Hozo, an ontology building tool strongly inspired by role-theory. The second one instead is about reusing in our model the conceptual primitives previously defined by one of the authors in the context of a more foundational type of work, the 'ontology of representations'.
We will briefly recap these two aspects in turns.

4.1 Role-based modeling with Hozo

Hozo[5] is a freely available ontology development environment created by Prof. Mizoguchi and his colleagues at Osaka University with the specific purpose of facilitating the creation of conceptual models that embed a correct distinction between role concepts and basic concepts (Mizoguchi et al. 2007).

---

4 It is important to stress that here we are not criticizing the use people make of this terms, for it is imprecise or inconsistent. Nor we are arguing against the characteristic ambiguity of human language. The main point we are making is that while humans deal surprisingly well with ambiguity in language, computers tend to fail. As a result, in a context such as the one defined by the semantic web initiative, using a language that is clear and unambiguous becomes a useful good practice, if not a necessity.

5 http://www.hozo.jp

At the heart of this modeling tool, there is the recognition that an in-depth understanding of roles is critical to ontology development. Quite often, role concepts (e.g.: *teacher*) are understood as types, even though they differ from so-called natural types (e.g.: *human*) because of their dynamic nature. Role theory, in general, analyses and attempts to provide a systematic account of the way role concepts are used in language (Steimann 2000). Although this is a particularly complex area, in recent years researchers have been progressing quite steadily (Halpin 2001; Masolo et al. 2004) (Renear 2007). The importance of roles for good conceptual modeling is now widely recognized, as it is recognized that improper modeling of roles will greatly influence reasoning and truth-maintenance along the *is-a* hierarchy that usually constitutes the taxonomical backbone of an ontology.

So, for example, while instances of basic concepts such as *human* cannot stop being instances of such concepts without 'dying' (disappearing from the real world), this is not the case with role concepts. For example, if we consider the *teacher* role, a human can stop being a teacher by just *leaving the role* without losing its identity. This is explained by one of the essential characteristics of roles, that is, roles are *played* by some entity (e.g., when we say that an instance of human is a *teacher* we mean that the teacher role is played by a human). However, it is impossible to find a player of human. Another important characteristic of roles is context dependence: this explains how and why an entity might change the role it plays according to a particular context. For example, a *man* would be regarded as a *teacher* in a *school* and as a *husband* in his *marital relationship*.

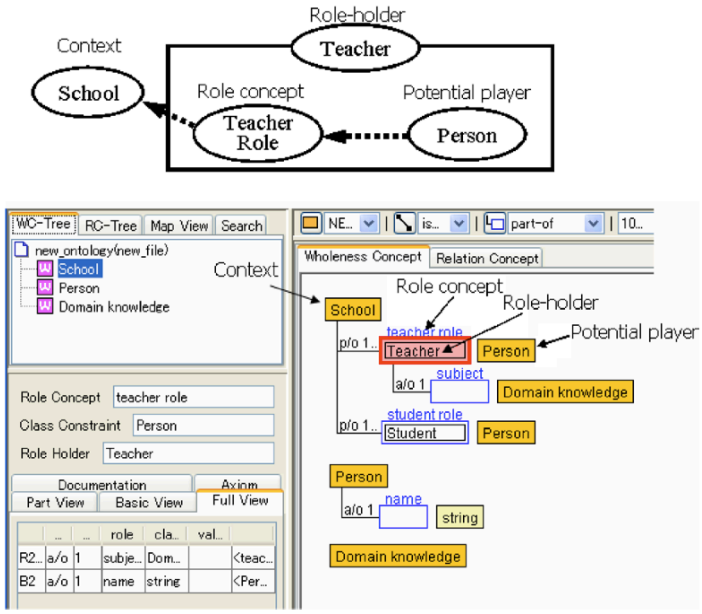

Fig. 3. Role-based modeling with the Hozo ontology engineering software

Hozo provides a highly visual, frame-based editing environment, that makes it simpler to create ontological models informed by role-theory. Hozo relies on a purpose-built knowledge representation language, but provides also import-export mechanism to other standard semantic web languages (such as RDF and OWL) so interoperability is guaranteed.

In Hozo each concept defined as a class is represented in a rectangle like *School* and *Person* (cf. fig. 3). Each class is defined by specifying its parts and/or attributes as slots. *School* is here defined as an entity composed of teachers and students where *teacher role* and *student role* are role concepts played by individuals specified by the rectangle at the far right, instances of *Person* in this case. In general, the basic philosophy behind Hozo's role model is that, in principle, all parts of a whole have their own roles to play in the context of the whole. For a more in depth description of the features of Hozo, we invite the reader to check (Mizoguchi et al. 2007) and (Sunagawa et al. 2006) .

4.2 Ontology of representations

The ontology of representations (Mizoguchi 2004) aims at providing the conceptual primitives needed to model the domain of 'content-bearing things', that is, very generally, entities that humans interpret as carrying some information (these objects have been called by other authors also 'information objects' (Gangemi et al. 2005) or 'information bearing objects' (Guha & Lenat 1990)).

A typical example is a *sentence*. What exist on the world wide web, that is, what we can be reached at URLs are not real entities but representations. Similar to that, there exist quite a few representations in the real world: *novel*, *poem*, *painting*, *music*, *procedure*, *symbol*, etc. What is the instance of a representation? How are representations different from real-world individuals? Can we 'copy' a representation? The ontology of representations provides a high level approach that allow answering to such questions via a formally defined ontological language.

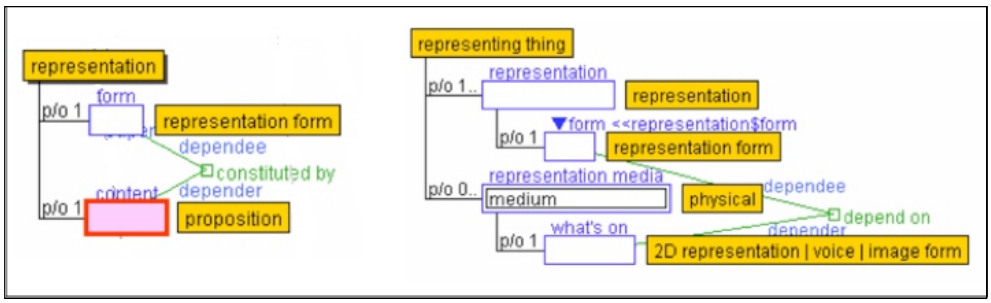

Fig 4. Ontological structure of representation and representing-thing.

For space reasons, we cannot expose all aspects of this theory. However, let us briefly outline two fundamental tenets behind it, for our next sections will be based on these ideas.

a) A representation is not concrete yet as it is, but it becomes a physical individual only when it becomes a representing thing. That is to say, there is a clear separation between *representation* and a *represented thing* (this approach can also be found on FRBR (IFLA & Saur 1997) and (Gangemi & Mika 2003)). The first one is a purely abstract (intangible) entity, while the second one is necessarily a concrete (tangible) one that contains it. Any representation is not embodied unless it becomes a representing thing. A sentence "This is a book" is a representation in the form of natural language (English) whose content is the meaning of "This is a book" and what you see is its printed realization on a sheet of paper which is a represented thing. Before writing it down, the symbol sequence "This is a pen" is not a physical individual because it does not specify what particular icon of symbols are used to represent each of the symbols. When the language is written down on a sheet of paper, it becomes a represented thing.

b) A representation is always composed of two parts, *form* and *content*. *Content* is the hidden part of representation, as it is the proposition the author of the representation would like to convey through the representation. On the other hand, *form* is the structure used to convey a meaning, such can be a symbolic language of some sort (e.g. a computer language or a musical score). Apparently, there can be many ways to code the same content into different forms so to produce different representations. Thus, the identity of "representation" is determined by the combination of both *form* and *content*. If their contents are different, then the representations are different. If instead two representations have the same content, their identity depends entirely on their form.

Keeping in mind the two points just introduced, we can now look at some typical modeling examples based on this ontology:

Algorithm representation is-a: Representation p/o"form": Language p/o"content": Algorithm *A representation of Quicksort algorithm with C* instance-of: Algorithm representation p/o"form": <C language> p/o"content": <Quicksort algorithm> *Musical score* is-a: Representation p/o"form": Musical symbol sequence p/o"content": a piece of music	*A score of symphony, the 5th* instance-of: Musical score p/o"form": <A sequence of musical symbols> p/o"content": <Symphony, the 5th> *A copy of a book of musical score of the 5th* instance-of: Representing thing p/o"representation": <A score of symphony, the 5th> p/o"medium": <Pieces of paper> *Sentence* is-a: representation p/o"form": natural language p/o: sequence of alphabets p/o"content": proposition(the meaning)

Fig. 5. Using the form/content pattern to model representations

5 Results: a role-based ontology for textual transmission history

The application of role-theory and of the ontology of representations allowed us to identify a number of modeling patterns. These have all been created using Hozo, and are here presented by means of snapshots taken from its user interface. In fig. 6 we can see a summary of the concepts we defined, within a (simplified) taxonomy.

It is important to remember that in this initial work we focused on the exploration of the applicability of these methods to the purpose of creating a precise semantic model for the EMLoT database. For this reason, we decided to concentrate our investigation on the conceptual aspects of the ontology and to leave aside, for the moment, the implementation details. In future work we will look at at how to best encode these results so to expose the resulting dataset using standard semantic web languages such as RDF and OWL (also thanks to Hozo's automatic translation functionalities).

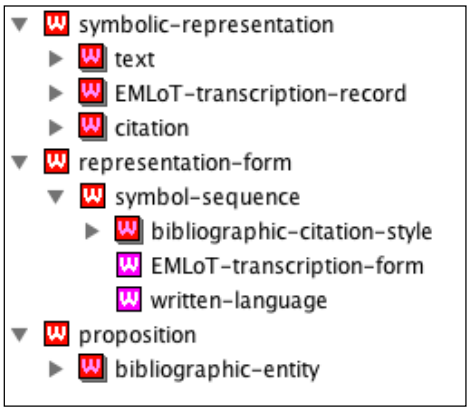

Fig. 6. Main ontology tree

5.1 Primary and Secondary sources.

The distinction between primary and secondary sources has fundamental importance in this project (as discussed in section 2, in EMLoT a primary source is by definition any document produced before 1642, while a secondary source is one produced after 1642). However it is also true that this distinction reflects the particular interests of the researchers's agenda, rather than some essential ontological feature sources possess in general. Therefore we concluded that all sources should be equally formalized by means of a subtype of `symbolic-representation`

called `text`, that is, the class of representations that are expressed using a representation form of type `written-language`. This modeling pattern is depicted in fig.7.

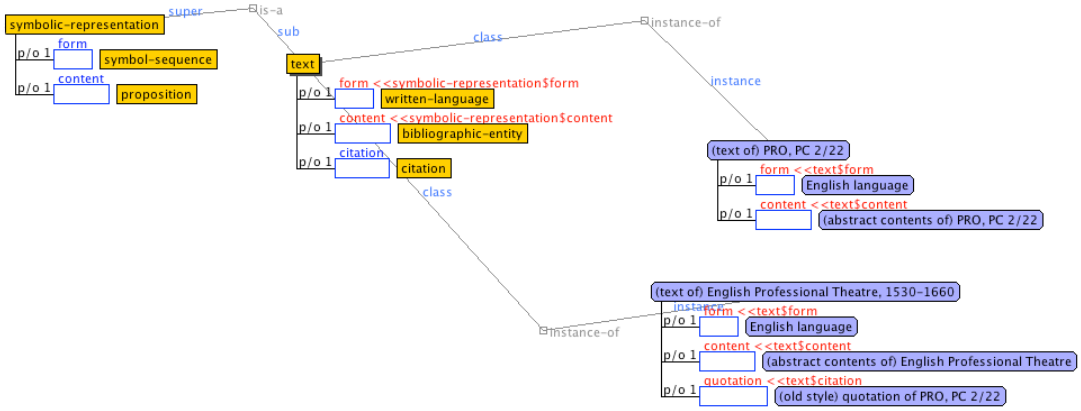

Fig 7. Modeling sources as symbolic representations

Instances of the class `text` can contain other representations too (eg a picture, a diagram, of fragments from another text); for the purposes of this explanation, in figure 7 we have highlighted only the fact that a `text` can contain one or more `citation` instance. For example in our scenario the "English Professional Theatre" instance contains a quotation from "PRO, PC 2/22" (other bibliographical information such as authors or titles here is omitted, although it would be defined at this level).

For that regards the primary and secondary status of these sources, as we will see in the next sections, these are best described as *roles* that a researcher assigns to them, in the context of a transcription record.

5.2 Reference, Citation and Quotation

The main intuition here is that while both citations and quotations are first-class entities, the notion of reference is better expressed in terms of a *role*: that is, the role played by a bibliographical entity in the context of a citation. According to our model, a `citation` (fig. 8) is to be intended as another specialization of `symbolic-representation` (fig. 7). What distinguishes it from other representations is its internal structure, that is, its *content* and *form*. The content of a citation is the abstract `bibliographical-entity` a citation refers to (hence we say that in this context the bibliographical entity plays the *reference* role). The form of a citation, instead, is the `citation-style`, that is, the class that formalizes at the abstract level the specific structure a citation uses in order to make reference to its object (cf. also section 5.3 below).

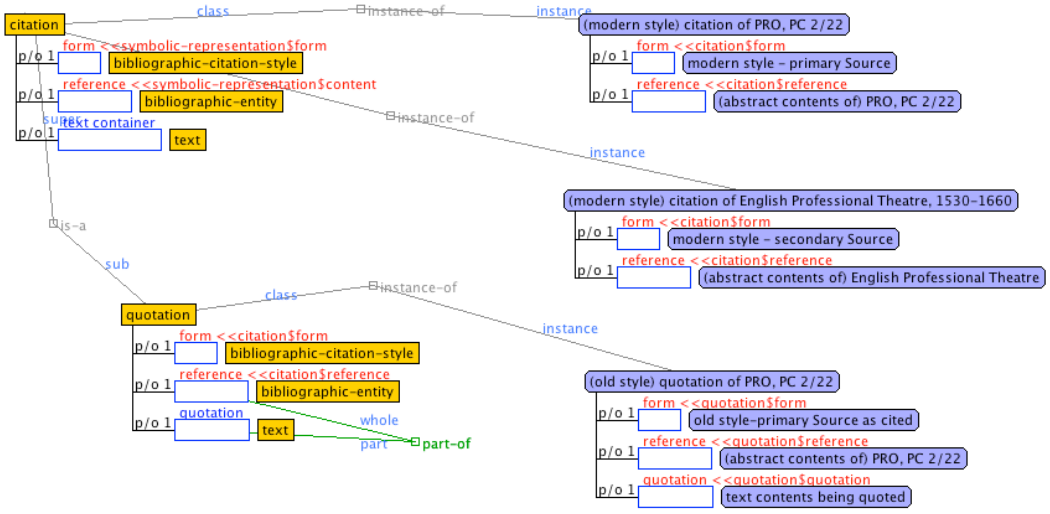

Fig 8. Modeling citations and quotations as symbolic representations

Finally, the conceptualization of quotations is very similar: a `quotation` is a subtype of `citation` that besides making reference to another bibliographical entity, includes also a textual fragment from it. In other words, we are arguing that every quotation is *also* a citation (although the opposite is not necessarily true). Please note that this modeling pattern requires a further logical constraint specifying that, in each `quotation` instance, the text being quoted *must* be part of the the bibliographic entity being referenced (in Hozo, this is expressed through a special *part-of* constraint, visible at the bottom left of fig. 8).

5.3 Citations styles

A `bibliographic-citation-style` is a type of `representation-form` (cf. fig. 9). It is the symbolic structure used to express the contents of a `citation`, to the purpose of highlighting certain aspects of a bibliographic entity, rather than others.

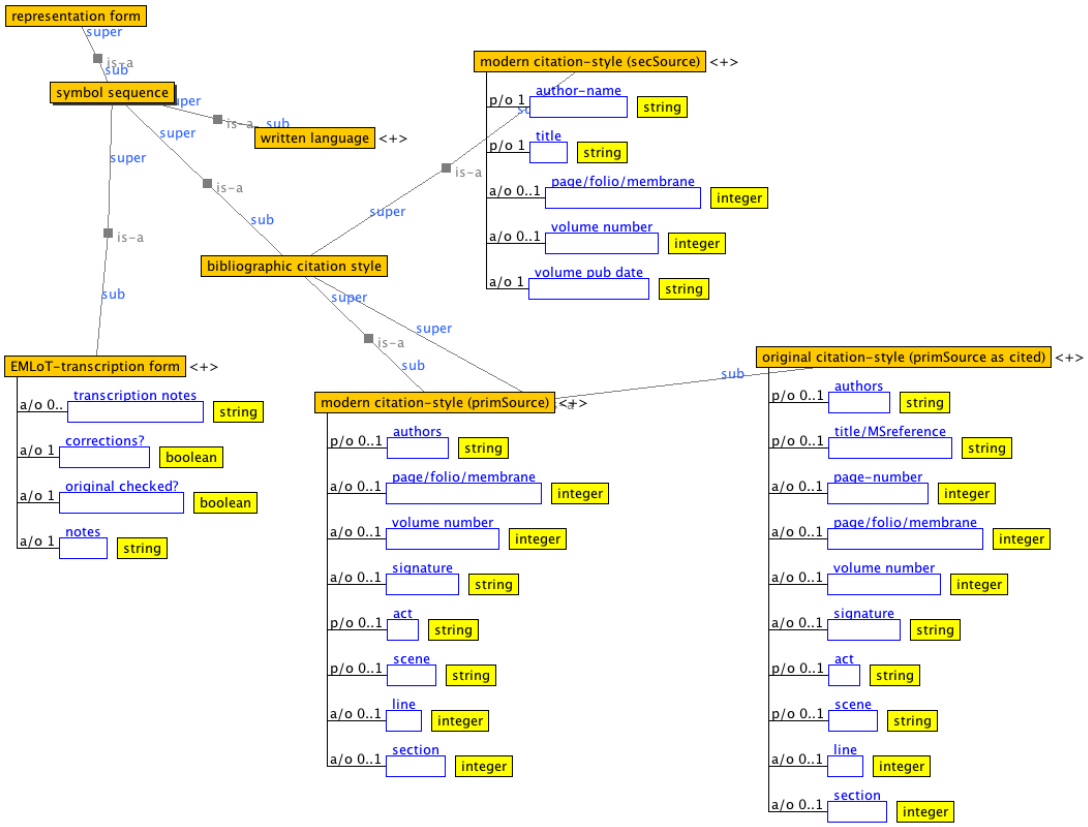

Fig 9. Bibliographic citation styles as representation forms

Normally a `bibliographic-citation-style` is composed by a number of atomic elements that correspond to some specific features of a document (eg author, title, publication place etc.); furthermore, it includes one or more rules determining how such elements need to be sequenced in order to construct a valid citation formula. Although it is not the main interest of this article, it is worth noting that this branch of the ontology could be developed further, so to model explicitly academic citation styles such as APA[6] or Chicago[7]. The resulting ontology could thus be used in order to automatically generate citations in different formats, or extract references from citations (as discussed for example in (Romanello & Pasin 2011)).

In EMLoT, a number of different citation styles have been employed: for example, within a transcription record we can find modern-style citations of both primary and secondary sources, and 'reconstructed' citation styles reflecting how the primary source was cited in the secondary source. As seen in figure 9, each one of these styles can be modeled as a separate entity that subclasses the `bibliographic-citation-style` representation form.

[6] http://www.apastyle.org/

[7] http://www.chicagomanualofstyle.org/

5.4 Transcription records

The last entity we would like to capture is the record generated by the EMLoT editors when they discover a transcription of a primary source within a secondary source. The structure of EMLoT transcription records reflects very specific curatorial and research interests (cf. section 2). Consequently, also its ontological counterpart should represent these meanings in a precise and transparent manner, for the simple fact that such records contains scholarly assertions that are themselves citable objects.

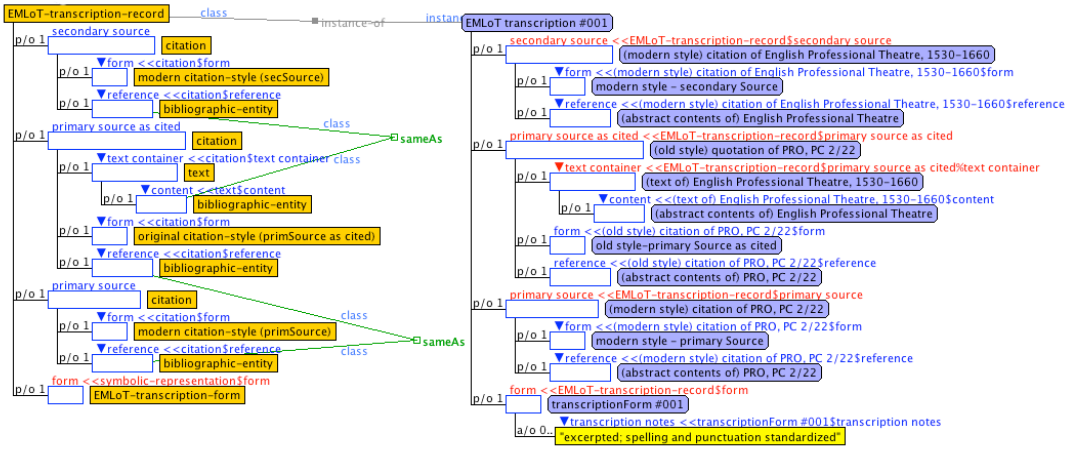

Fig 10. Expressing the semantic content of EMLoT transcription records

We decided to formalize transcription records by creating an ad-hoc subclass of `symbolic-representation`, called `EMLoT-transcription-record`. Figure 10 shows the structure of an `EMLoT-transcription-record`: two `citation` instances are given the roles of *primary* and *secondary* sources in the context of each single 'record', and are rendered via modern-style representation forms. Moreover, an `EMLoT-transcription-record` includes also the *primary-source-as-cited* role, which is played by another `citation` instance. In particular, this third citation object is actually just another reference to the primary source text, although this time the reference is expressed using a non-modern citation style, that is the `original-citation-style` representation form.

As a result, an `EMLoT-transcription-record` instance contains three citations: two modern ones, and an 'old' one reflecting the way the primary source was originally quoted in the secondary source (please note that since in our model a `quotation` is defined as a sub-type of

`citation`, both objects can play the roles of *primary* and *secondary* sources within an `EMLoT-transcription-record` instance).

Finally, the ontological schema includes also two important *same-as* constraints: the first one specifies that the primary source cited in the `EMLoT-transcription-record` (via the `modern-citation-style`) and the one transcribed in the secondary source (via the `original-citation-style`) must be the same. The second one instead is required so to make sure that the secondary source text and the text where the quotation was found, are indeed the same. A specific subtype of `representation-form` called `EMLoT-transcription-form` (cf. also fig. 9 above) provides the logical structure necessary for encoding the key differences between the original and the transcription, as highlighted by the editors (in the case shown, we only have some transcription notes, i.e. "excerpted; spelling and punctuation standardized").

This concludes our analysis, for all the textual entities introduced at the beginning of this essay (and depicted in fig. 1 above) have been re-conceptualized in a more generic and application independent fashion.

6 Conclusion and future work

In this paper we have presented an approach based on role theory and ontological engineering aimed at the formal characterization of common concepts in the discipline of textual transmission history. Our analysis took part from the EMLoT project, the first phase of which completed in 2011 with an online database providing information of what use of pre-1642 theatrical documents has been made in post-1642 materials.

The database highlighted a number of conceptual modeling issues that derive from an in-depth analysis of the way texts cite each other in the humanities, and how humanities scholars tend to study and analyze this networks. The ontological solutions provided in the article facilitate the representation of such concepts using a more transparent semantic language; as a result, the contents of the database could be expressed using this language and made available on the semantic web so to be integrated with other resources.

The ontological analysis herewith presented showed also that the task of modeling citations (and more broadly, bibliographical data) in the humanities brings forward scenarios that are not entirely reducible to corresponding work done in more scientific domains. The key insight here is that the very objects of interest in humanities scholarly activities (such can be the ones of textual transmission history) reveal a quite complex and partially different nature than their

scientific counterparts, for they are often rooted in radically different contexts. As a result, we hope to have convinced the reader that to the purpose of creating a truly useful web of data infrastructure in the digital humanities, it is mandatory to avoid reusing pre-packaged solutions coming from other domains; rather, we should expect to find objects that require a different analysis altogether.

We are currently working towards a working implementation of the solutions discussed in this article, using the standard language made available by the semantic web research community. Once that is done, the EMLoT database will be providing a dedicated access point for semantic web softwares, so to facilitate the integration of its data with other resources online.

Acknowledgments. This work has been carried out thanks to the generous funding of the AHRC, the SSHRC, and the Japanese Society for the Promotion of Science.

References

Auer, S. et al., 2007. DBpedia: A Nucleus for a Web of Open Data. In 6th International Semantic Web Conference (ISWC 2007).

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American*.

Brickley, D. & Miller, L., 2010. FOAF Vocabulary Specification. *Namespace Document*, 3 (Revision 1.113). Available at: http://xmlns.com/foaf/spec/ [Accessed April 22, 2012].

Cover, R., 1998. XML and Semantic Transparency. *xml.coverpages.org*, pp.1–4. Available at: http://xml.coverpages.org/xmlAndSemantics.html [Accessed April 22, 2012].

Crofts, N. et al., 2008. *CIDOC-CRM Version 4.2.4 - Reference Document*. Available at http://www.cidoc-crm.org/previous_releases_cidoc.html [Accessed April 22, 2012].

D'Arcus, B. & Giasson, F., 2009. Bibliographic Ontology Specification | The Bibliographic Ontology. Available at http://bibliontology.com/specification [Accessed April 22, 2012]

Davis, R., Shrobe, H. & Szolovits, P., 1993. What is a Knowledge Representation? *AI Magazine*, 14(1), pp.17–33.

Doerr, M., 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine archive*, 24(3), pp.75–92.

Gangemi, A. & Mika, P., 2003. Understanding the Semantic Web through Descriptions and Situations. In International Conference on Ontologies, Databases and Applications of Semantics (ODBASE).

Gangemi, A., Borgo, S. & Catenacci, C., 2005. *Metokis deliverable D07 - Task Taxonomies for Knowledge Content*,

Garshol, L.-M., 2004. Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all. *Journal of Information Science*, (30), pp.378–391.

Gruber, T., 1992. The Bibliographic Data Ontology. *Technical Report*. Available at http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/bibliographic-data/ [Accessed April 22, 2012]

Gruber, T., 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies - Special issue: the role of formal ontology in the information technology*, 43(5-6).

Guha, R. & Lenat, D.B., 1990. Cyc: A Midterm Report. *AI Magazine*, pp.1–28.

Halpin, T., 2001. Object Role Modeling: An Overview. *Microsoft Technical Report*, pp.1–16.

Heath, T. & Bizer, C., 2011. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1), pp.1–136.

IFLA & Saur, K., 1997. Functional requirements for bibliographic records: final report (amended on Feb 2009). *International Federation of Library Associations and Institutions*.

Kruk, S., Synak, M. & Zimmermann, K., 2005. MarcOnt: integration ontology for bibliographic description formats. *DCMI '05: Proceedings of the 2005 international conference on Dublin Core and metadata applications: vocabularies in practice*.

MacLean, S.-B., Hagen, T. & Pasin, M., 2012. Moving Early Modern Theatre Online: the Records of Early English Drama Introduces the Early Modern London Theatres Website. In T. Gniady, ed. *New Technologies in Medieval and Renaissance Studies*. Medieval and Renaissance Texts and Studies Series (Iter Inc./ACMRS).

Masolo, C. et al., 2004. Social Roles and their Descriptions. In Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR2004), pp. 267-277.

Mizoguchi, R., 2004. Tutorial on ontological engineering - Part 3: Advanced course of ontological engineering. *New Generation Computing*, 22(2), pp.198–220.

Mizoguchi, R. et al., 2007. A Model of Roles within an Ontology Development Tool: Hozo. *J. of Applied Ontology*, pp. 159–179.

Palmer, S.B., 2001. The Semantic Web: An Introduction. *infomesh.net*. Available at: http://infomesh.net/2001/swintro/ [Accessed September 14, 2011].

Poli, R. & Obrst, L., 2009. The Interplay Between Ontology as Categorial Analysis and Ontology as Technology. *TAO-Theory and Applications of Ontology*, pp.1–25.

Portal, A., 2002. Advanced Knowledge Technologies - Portal Ontology. *www.aktors.org*. Available at: http://www.aktors.org/publications/ontology/ [Accessed June 2, 2009].

Renear, A.H., 2007. Three of the Four FRBR Group 1 Entity Types are Roles, not Types. In 70th Annual Meeting of the American Society for Information Science and Technology (ASIST).

Romanello, M. & Pasin, M., 2011. An Ontological View of Canonical Citations. In Digital Humanities 2011.

Shotton, D., 2009. CiTO, the Citation Typing Ontology, and its use for annotation of reference lists and visualization of citation networks. In Bio-Ontologies 2009, a Special Interest Group meeting at ISMB 2009.

Steimann, F., 2000. On the Representation of Roles in Object-Oriented and Conceptual Modelling. *Data & Knowledge Engineering*, 35, pp.83–106.

Sunagawa, E. et al., 2006. Role Organization Model in Hozo. *Managing Knowledge in a World of Networks*, pp.67–81.

Uschold, M., 2003. Where are the Semantics in the Semantic Web? *AI Magazine*, 24(3).