A computational analysis of art historical linked data for assessing authoritativeness of attributions

A Computational Analysis of Art Historical Linked Data for Assessing Authoritativeness of Attributions

Marilena Daquino
University of Bologna marilena.daquino2@unibo.it

Abstract

In this paper a comparative analysis of art historical linked open data is presented. The result of the analysis is a conceptual framework of Information Quality measures designed for validating contradictory sources of attribution on the basis of a documentary, evidence-based approach. The aim is to develop an ontology-based ranking model for recommending artwork attributions and support historians and cataloguers' decision-making process. The conceptual framework has been evaluated by means of a user study and the evaluation of a web application leveraging the aforementioned ranking model. Results of the survey demonstrate that findings satisfy users' expectations and that are potentially applicable to other types of information in the Arts and Humanities field.

Keywords: Information quality, authoritativeness, data mining, Linked Open Data, art history

Introduction

Attributions made in the Arts field are characterized by a high degree of uncertainty and questionability. In particular, *connoisseurs* ascribe artworks to artists on the basis of their knowledge on fine arts. Despite several scientific methodologies have been proposed for validating artwork attributions (Morelli & Richter, 1883; Ginzburg, 1979; Freedberg, 2006), these are not reproducible, hence many attributions are still debated. Secondly, cultural institutions rely on primary and secondary sources when supporting artwork attributions recorded in their catalogues, including articles, books, auction and museum catalogues, and the aforementioned scholars' opinions. Cataloguers review contradictory attributions and select the most authoritative ones according to a number of criteria that are deemed reliable in the community (e.g. bibliography on the topic, stylistic analysis). Similarly, scholars consult online catalogues and gather documentation for comparative studies (Brilliant, 1988) so as to validate attributions. They evaluate a number of context information, such as the number of trusted institutions in agreement on a certain attribution, the reputation of historians that first ascribed the artwork, whether information is recorded in updated, scholarly, and peer-reviewed evidences, and so on.

Since methodologies for validating attributions are not reproducible, authoritativeness of sources is a key aspect in the Arts domain (Freedberg, 2006). However, a formal definition of authoritativeness is still an open issue, that regards both primary sources of attribution (connois-

seurs) and secondary sources (cataloguing records, multipurpose websites). Moreover, gathering secondary sources and evaluating aspects characterizing their authoritativeness is an expensive and time-consuming task, completely demanded to the user and easily error-prone. Biased information may be recorded (due to market interests), sources might not be updated or may include partial information (e.g. not documented attributions).

Aggregators of art historical data, e.g. Europeana[1] and Pharos[2], show an increasing interest in supporting users in common tasks such as gathering heterogeneous sources of information. Nonetheless, existing services do not explicitly handle contradictory information, nor they support the assessment of reliability. Therefore, sophisticated questions such as "what is the most documented and shared attribution for the artwork at hand?" are demanded to the user's subjective analysis.

In this paper we argue that quantitative methods and Semantic Web technologies can support users' decision-making process when gathering and reviewing online secondary sources recording artwork attributions. By relying on a documentary, evidence-based approach to appraise secondary sources, we aim at formally defining features characterising authoritativeness of information sources in the Arts field, that is, their *textual authority*.

In order to identify Information Quality (IQ) dimensions that characterize textual authority, we created a corpus of cataloguing records describing artworks provided by three representative art historical photo archives, namely the Federico Zeri Foundation[3], Villa I Tatti - Berenson Library[4], and the Frick Art Reference Library[5]. Records include detailed information on argumentations around attributions that are either accepted or discarded by the cataloguing institution. Cataloguing metadata is transformed into a common representational model, i.e. RDF[6], and the resulting dataset is queried for knowledge extraction purposes.

In particular, we obtained (a) a controlled vocabulary of terms (also *criteria* from now on) that represent motivations supporting an attribution (e.g. bibliography, scholar's attribution, museum attribution), (b) a rating of such criteria validated by domain experts and data analysis, (c) a number of other IQ metrics for measuring textual authoritativeness, (d) a preliminary work on metrics for defining scholars' authority, and (e) a ranking model for measuring authoritativeness of sources of attribution. So doing, we aim at reproducing cataloguers and art historians' hermeneutical approach when validating contradictory attributions and generalize findings so that they can be applied to near fields or similar information.

The paper is structured as follows. In section *Related Work* previous work on IQ measures and authoritativeness is introduced. In *Problems, Questions, and Limitations*, research questions, assumptions and restrictions for defining authoritativeness are addressed. In *Research study* we describe the research design, the corpus analysis, the conceptual framework of IQ measures and the resulting ranking model. In section *User Study*, we discuss the user study performed to evaluate the soundness of the conceptual framework. Lastly, in *Results and Discussion* we discuss findings and limits of our approach, and we conclude with new research lines for future work.

Related Work

Information Quality is the fitness for purpose of information, which encompasses several domain-dependent and independent dimensions. Research fields address IQ dimensions differently. In Library and Information Science, scholars and librarians developed guidelines and checklists (Cooke, 1999) and focused on functional aspects of metadata (Park, 2009). Cataloguing and metadata standards (Moro, Mancinelli, & Negri, 2017; Baca & Harpring, 2006; McKenna & Patsatzi,

2007; Coburn, Light, McKenna, Stein, & Vitzthum, 2010), vocabularies (Doerr, 2009, 2003; Peroni & Shotton, 2018; Daquino, Mambelli, Peroni, Tomasi, & Vitali, 2017), and thesauri (Baca & Gill, 2015) naturally cover aspects peculiar of the Arts field. However, how to support stakeholders in assessing reliability of questionable information is not taken into account. So far methods for modelling and reasoning on argumentation (Walton, 2013) and reliability of statements, haven't been considered neither in cataloguing practices, nor in the Arts field.

Computer Scientists developed frameworks and methodologies for data quality assessment (Lee, Strong, Kahn, & Wang, 2002; Batini, Cappiello, Francalanci, & Maurino, 2009). (Knight & Burn, 2005) reviewed the most common dimensions available in a number of IQ frameworks. (Naumann & Rolker, 2005) defined a set of IQ dimensions and a three-fold assessment approach, namely: (a) Subject criteria (the user); (b) Object criteria (the information source); (c) Process criteria (the information retrieval process). Semantic Web technologies have been widely used for tracing and representing data provenance so as to assess trustworthiness of statements (Moreau, Groth, Cheney, Lebo, & Miles, 2015; Zaveri et al., 2016), but no work exists on the formal definition of authoritativeness in the Arts and Humanities field.

The definition of authoritativeness has been addressed in several works. According to Wilson (Wilson, 1983), *cognitive authority* refers to the extent to which a second-hand information provider is deemed trustworthy. This applies to cultural institutions publishing cataloguing records including artwork attributions. (Rieh, 2002) focuses on users' judgment, and includes authoritativeness in the list of dimensions characterizing cognitive authority, namely: trustworthiness, reliability, scholarliness, credibility, officialness, and authoritativeness. (Farahat, Chen, Mathis, & Nunberg, 2007) analyzed types of authority that affect information retrieval tasks. *Social authority* is a graph-theoretical notion that can be measured by relying on social networks, e.g. citation indexes, lists of trusted providers. Secondly, they introduce *textual authority*, a non-topical estimate of the intrinsic quality of a source, that is, the extent to which information is useful, good, current, and accurate.

In this study we present the analysis of a corpus of contradictory artwork attributions recorded in cataloguing records. We rely on a subset of IQ assessment methods proposed by (Naumann & Rolker, 2005) so as to measure IQ dimensions characterizing textual authority of art historical data sources and recommend authoritative attributions. We also propose a preliminary work on metrics for assessing cognitive authority of scholars cited as sources of attribution.

Problems, Questions, and Limitations: Defining Authoritativeness in the Arts Field

When recording attributions, cultural institutions do not explicit which methods were used to validate an attribution. Criteria to be adopted are listed in cataloguing standards, but there is no guidance on how to rate (and weight) such criteria. Cataloguers may adopt different criteria, characterized by different degrees of reliability, according to the context. For instance, a cataloguer may discard an attribution that is claimed by an auction firm and prefer an attribution recorded in a peer-reviewed article. However, a recent discovery made by another auction firm can overtake the outdated article. Despite a shareable rating of criteria should address an *a priori* approach for reviewing contradictory statements, several factors can affect the choice and should be taken into account too.

The proposed research aims at providing theoretical foundations and technical solutions for assessing attributions in online secondary sources. In particular, the study focuses on the formalization of the dimensions characterizing the hermeneutical approach of art historical data providers

and users' judgment. The following research questions are addressed: (a) What are the criteria characterizing the hermeneutical approach of art historical data providers when reviewing attributions? (b) Can we address and measure dimensions characterizing authoritativeness of secondary sources recording attributions? (c) Can we address and measure features characterizing authoritativeness of scholars that first ascribed an artwork?

In order to answer such questions we rely on the design-science method proposed by (Hevner, March, Park, & Ram, 2004), which seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts. In this research, an artifact for harvesting and consuming art historical data and for supporting users' decision-making process is developed. The artifact leveraging findings here presented is called *mAuth - mining authoritativeness in art history*[7].

Beneficiaries of this research are several. Art historians, cataloguers, and art dealers can benefit of specialized applications for gathering sources of attribution and save time. Second, aggregators can highlight immediately reliable and well-documented attributions and enable users to compare different scientific approaches adopted by data providers, showing (eventually) whether information is biased (e.g. attributions made by art dealers or auction firms). Lastly, bespoke policies and services for metadata quality improvement can leverage our findings for automatically updating poor-quality, older metadata and avoid time-consuming and expensive tasks.

As aforementioned, the study focuses on artwork attributions, but results can be applied *mutatis mutandi* to similar types of information. We focus on the appraisal of secondary sources only, while the judgment of the artwork itself is demanded to art historical data providers. To this extent, we rely on Wilson's definition of cognitive authority of second-hand knowledge providers. Secondly, we narrow the comparative analysis to cataloguing records provided by art historical photo archives. Photo archives used to be research places for connoisseurs, hence these are likely to preserve insights on contradictory attributions. On the contrary, museum and gallery records are excluded from the analysis since these do not offer the same insights. Lastly, metrics proposed for measuring cognitive authority are in a very early stage, due to the lack of representative databases providing historical citation data for the Arts and Humanities field.

Research Study

The study can be divided in three phases, namely (a) corpus analysis, (b) definition of IQ measures, and (c) development of a ranking model.

In summary, seven actions (S1 to S7) were undertaken in order to achieve the final ranking model. First, content standards are reviewed so as to extract an initial set of terms identifying criteria that cataloguers are allowed to use in records when motivating an attribution (S1). Terms from the controlled vocabulary are reconciled to descriptive fields including attributions in the corpus, and the original vocabulary is pruned and refined (S2). The resulting set of terms is revised by domain experts and a first rating of those is provided by using a 1-10 scale (S3). The rating is validated by analysing whether terms are consistently used in the corpus for supporting accepted or discarded attributions according to the rating (S4). Other IQ dimensions affecting the reliability of an attribution are selected from prior works (S5) and bespoke metrics are developed (S6). Finally, IQ measures are weighted and combined in the ranking model (S7).

Corpus analysis

The objective of the corpus analysis is to define the most shareable rating of criteria adopted by cataloguers when recording attributions in cataloguing records.

The corpus is gathered on a topic base, i.e. attributions of artworks of the Modern Era, and includes three datasets. In detail, the Federico Zeri photo archive contributed with 19.061 records, Villa I Tatti with 12.256 records, and the Frick Art Reference Library with 10.207 records. Records include argumentations around attributions in the form of discursive text fields, which can be reconciled to (one or more) terms identifying criteria. For instance "Federico Zeri's attribution (1979)" can be classified as a scholar's attribution, "Christie's attribution (1928)" as an auction attribution. It's worth to notice that records may include both attributions accepted by the cataloguing institution and discarded attributions, recorded for historical reasons. For instance, a record may include the following statement: "Attribution: Andrea Verrocchio, Federico Zeri's attribution (1979). Other attributions: Leonardo da Vinci, Christie's attribution (1928)".

The analysis is performed over the Linked Data version of the three datasets[8] rather than the original XML collections since (a) data cleansing and data reconciliation techniques have been applied to the RDF dataset, and (b) the semantic interoperability makes easier the comparative analysis.

S1. Review of Cataloguing Standards and Guidelines. Content standards and guidelines for cataloguing artworks detailed in Section *Related Work* include lists of terms identifying criteria that cataloguers can use to specify the main reason for supporting an attribution. We collect such terms in order to address the broadest scope of our scenario. The ICCD-OA standard (Moro et al., 2017) resulted being the most comprehensive controlled vocabulary, including nineteen terms, namely: diagnostic measures, iconographic analysis, stylistic analysis, historical analysis, type analysis, bibliography, stamp, mark, inscription, archival classification, comparison, context, documentation, artist's analysis, handwriting style, signature, monogram/sigla, handwritten note, traditional attribution.

S2. Refinement of the controlled vocabulary. We reconciled discursive argumentations around attributions included in the corpus to (a) linked data entities representing people (scholars) and organizations (museums, galleries, auction firms) by using a number of data reconciliation methods, and (b) terms belonging to the aforementioned controlled vocabulary by using regular expressions. The objective of the reconciliation is to address which terms are currently adopted in three representative scenarios.

The analysis shows that (a) the initial controlled vocabulary includes a number of terms that are not used in the corpus of records, hence not all terms can be evaluated, while only nine criteria out of nineteen are actually used, and (b) nine new criteria (not included in any prior standards) were found. For the sake of simplicity we reduce some terms under the same definition, e.g. mark and inscription into inscription, and we add the fuzzy terms *other* and *none* for labelling argumentations that do not fall into any classification. The result is again a list of nineteen criteria, namely: documentation, artist's signature, bibliography, archival classification, scholar's attribution, museum attribution, scholar's note on photograph, inscription, sigla, auction attribution, collection attribution, market attribution, traditional attribution, stylistic analysis, anonymous note on photograph, false signature, caption on photograph, other, none.

S3. Domain experts' revision and first rating of criteria. Cataloguers with a background in art history from the Federico Zeri photo archive were asked to double-check the list of nineteen criteria and to provide a first rating of those by using a 1-10 scale (where 1 is the less authoritative

and 10 is the most authoritative criterion). The objective is to achieve a first rating on the basis of domain experts' consultancy.

We notice that cataloguers tend to prefer attributions provided by scholarly authorities and attributions derived from the appraisal of photographic documentation. The resulting list of criteria (with the associated score in parentheses) is the following: documentation (10), artist's signature (9), bibliography (8), archival classification (7), scholar's attribution (6), museum attribution (5), scholar's note on photograph (5), inscription (5), sigla (5), auction attribution (4), collection attribution (4), market attribution (4), traditional attribution (4), stylistic analysis (3), anonymous note on photograph (3), false signature (2), caption on photograph (2), other (1), none (1).

S4. Validation of the rating over the three datasets. The rating proposed by domain experts is validated by checking its consistency over the three datasets. In particular, given a subset of records including both accepted attributions and discarded attributions for the same artwork, criteria that support accepted attributions are compared one-by-one to criteria that support discarded attributions. So doing we want to quantify whether criteria that are supposed to be less/more reliable in the rating are consistently used or not. For instance, how many times scholars' attributions are consistently deemed more reliable than auctions' attributions (according to the rating) for the same artwork? The result is a 19x19 table where all the criteria are (potentially) compared with each other. The aim is to confirm or revise domain experts' rating and highlight whether criteria can be rated *a priori* (i.e. these are always valid) or other factors may have affected the final decision of cataloguers (highlighted by inconsistencies in data).

The subset here analysed includes 5.356 records from Zeri, 5.384 from Villa I Tatti, and 941 from Frick. Data from Villa I Tatti and Frick were not published before as linked data and were provided as .csv files. Tabular data were transformed into RDF according to the same ontologies already used by the Zeri photo archive. We first analyse the three subsets individually and secondly we merge data to have a broader overview.

The Federico Zeri photo archive. Figure 1 shows the distribution of paired criteria in the Zeri dataset. Rows represent criteria supporting all the accepted attributions and columns represent criteria supporting discarded attributions for the same artworks. Cells at the intersection between columns and rows represent the number of times the criterion supporting an accepted attribution (i.e the value in the row) is preferred over the criterion supporting a discarded attribution (i.e the value in the column) for the same artwork. More than one criterion may support an attribution, hence there is an overlap in the usage of criteria. Values in columns "tot." represent the total number of records that use the criterion at hand. For instance, "documentation" supports an accepted attribution in 34 records; 21 times out of 34 it is preferred over a discarded attribution that is motivated by a "scholar's attribution" (first row, third column). However, attributions supported by "documentation" are discarded 69 times when another attribution is supported by a "scholar's attribution" (third row, first column). Empty cells represent criteria that are never compared in contradictory attributions for the same artwork.

The distribution shows that the archival classification of photographs depicting artworks supports the 99% attributions (i.e. 5.322 records) at the Zeri Foundation. Secondly, scholars' attributions (2629, i.e., 49%), and bibliographic references (1697, i.e., 32%) are the main tools for validating attributions. Some criteria are not well represented in the dataset, such as museums attributions, collection attributions and traditional attributions. In such cases we trust the original rating provided by experts.

We notice some inconsistencies in the usage of criteria. According to archivists, "documenta-

tot.	ACCEPTED \ DISCARDED	tot.	documentation	artist's signature	scholar's attribution	bibliography	archival classification	scholar's note on photo	museum attribution	inscription	sigla	auction attribution	collection attribution	market attribution	traditional attribution	stylistic analysis	anonymous note on photo	false signature	caption on photo	other	none
		108	8	783	2547	334	318	88	6	4	701	153	227	8	1	1259	11	56	4	0
34	documentation		15		21	12	13	3			3		4			4				
26	artist's signature			3	5	13	2	3			1	1				3				
2629	scholar's attribution		69	3	527	973	194	144	42	4	3	436	76	142	2	1	798	6	24	3	
1697	bibliography		17	2	253	1288	199	81	29			96	32	16	5		201		24	1	
5322	archival classification		108	8	795	2585	328	315	88	6	4	700	153	226	8	1	1246	11	56	4	
471	scholar's note on photo		6	1	49	218	53	110	5	1		58	12	21			102		6		
1	museum attribution				1		1		1												
3	inscription					3															
2	sigla					1											1				
73	auction attribution				35	17	39	8				48	1	4			3				
13	collection attribution				5	1	6	3					13				1				
28	market attribution				13	5	11					2	1	13			3				
0	traditional attribution																				
8	stylistic analysis				1	1		1				4					1	1			
110	anonymous note on photo				43	37	56	8				8	2	1			72		1		
0	false signature																				
5	caption on photo				4	1	5												5		
111	other		8		9	26			1			16	4	7			46				
132	none		6		14	45		4	2			17	6	5			45	1	2		

Figure 1. Distribution and comparison of criteria adopted by the Zeri photo archive

tion" (i.e. expertises documented by art historians) is deemed the most reliable criterion. However, it is discarded when the accepted attribution is supported by "archival classification" (108), scholar's attribution (69), bibliography (17), and scholars' notes on photographs (6). Further analysing the latter criteria we notice that (1) 64 out of 69 are Federico Zeri's attributions (i.e. archive creator's attribution), (2) 1 out of 17 is Zeri's bibliography (i.e. archive creator's bibliography), and (3) 6 out 6 are Zeri's annotations (i.e. archive creator's note on photograph). A similar inconsistency is found between "scholar's note on photo" and "bibliography". The former is preferred 218 times over the second and discarded 81 times. We notice that 81 annotations out of 81 are signed by Federico Zeri. We deduce that cataloguers are biased by the archive creator's opinion. To this extent, it is worth to notice that (a) 2513 out of 2641 scholars' attributions are Federico Zeri's attributions, (b) 169 out of 1714 bibliographic references are Federico Zeri's publications, and (c) 471 out of 471 annotations on photographs are made by Federico Zeri.

The Villa I Tatti photo archive. Figure 2 shows the distribution of criteria in the dataset provided by Villa I Tatti - Bernard Berenson Library. The archive is pretty similar to the Zeri photo archive, that is, these are both created by art historians, they often describe same artworks, and the methodology to assess attributions is likely to be similar, or comparable. Therefore, we include here the three situations highlighted in the Zeri dataset, namely archive creator's bibliography, archive creator's attribution and archive creator's note on photograph.

tot.	ACCEPTED / DISCARDED	documentation	artist's signature	archive creator's attrib.	archive creator's bibl.	bibliography	archival classification	archive creator's note	scholar's attribution	scholar's note on photo	museum attribution	inscription	sigla	auction attribution	collection attribution	market attribution	traditional attribution	stylistic analysis	anonymous note on photo	false signature	caption on photo	other	none
	tot.	39	0	20	279	290	4	710	483	396	19	23	0	202	14	1	0	0	422		8	7	0
3	**documentation**					1		1	1										1				
12	**artist's signature**							8	1			1							3				
1	**archival creator's attrib.**																		1				
368	**archival creator's bibl.**	1		3	72	118		130	93	53	2	3		27	5				52		2	3	
284	**bibliography**	3		6	133	36		111	48	39	1			19	3				14			1	
680	**archival classification**	17		4	8	92		185	177	131	3	13		73	3	1			184		2	2	
514	**archival creator's note**	11		6	3	36	1	256	92	110	11	2		44	5				104		1	1	
61	**scholar's attribution**				11	10	2	18	6	15				5					15			1	
194	**scholar's note on photo**	5		1	3	17	1	59	51	70	1	1		22	2				37		1		
93	**museum attribution**			2	29	4		50	19	25		1		3					9		1		
0	**inscription**																						
0	**sigla**																						
95	**auction attribution**	1		1	13	3		42	17	12		3		15	1				13				
11	**collection attribution**				5			7						2					2				
0	**market attribution**																						
0	**traditional attribution**																						
0	**stylistic analysis**																						
63	**anonymous note on photo**	1			2	4		13	19	11	1			7					27		1		
0	**false signature**																						
0	**caption on photo**																						
0	**other**																						
0	**none**																						

Figure 2. Distribution and comparison of criteria adopted at Villa I Tatti

Like in the Zeri photo archive, preferences reflect peculiarities of the photo archive, namely: (a) an extensive usage of "archival classification", (b) the influence of scholars' opinions and annotations, and (c) the usage of bibliography, specifically Berenson's references. Likewise, the archive creator's opinion appears more reliable than other scholars' attributions and notes on photographs. Other criteria do not provide insights on the actual preferences since these are either underrepresented or completely absent. The core of criteria characterizing the methodology seems to be shared between Zeri and I Tatti. The actual usage of criteria confirms the prior rating, but no further information can be deduced on other criteria.

The Frick Art Reference Library photo archive. Figure 3 shows the distribution of criteria in the dataset provided by the Frick Art Reference Library. The dataset has been chosen to validate the rating over a different type of photo archive, that is, an archive that is not created by a scholar. Therefore, the methodology does not include references to a predominant scholar, such as "archive creator's attribution".

Some criteria are underrepresented. However, similarities in the usage of highly rated criteria

tot.	ACCEPTED \ DISCARDED	documentation	artist's signature	archive creator's bibl.	archive creator's attrib.	bibliography	archival classification	archive creator's note	scholar's attribution	scholar's note on photo	museum attribution	inscription	sigla	auction attribution	collection attribution	market attribution	traditional attribution	stylistic analysis	anonymous note on photo	false signature	caption on photo	other	none
	tot.	1	1	0	0	121	123	0	561	18	73	2	0	40	12	0	7	0	39	0	1	3	115
1	**documentation**	1																					
0	**artist's signature**																						
0	**archival creator's attrib.**																						
0	**archival creator's bibl.**																						
434	**bibliography**					25	13		298	2	32				5				15				44
375	**archival classification**		1			41	60		142	6	24	2		25	2		3		7			2	60
0	**archival creator's note**																						
167	**scholar's attribution**					18	29		86	2	7			7	3		2		9			1	3
7	**scholar's note on photo**					1			3	3													
37	**museum attribution**						12		11		3	1		1	1		2						6
0	**inscription**																						
0	**sigla**																						
6	**auction attribution**													2	1				1				2
4	**collection attribution**						1				2			1									
0	**market attribution**																						
0	**traditional attribution**																						
15	**stylistic analysis**					2			7	4				2									
69	**anonymous note on photo**					34	6		13	1	5			2	1				7				
0	**false signature**																						
0	**caption on photo**																						
2	**other**						2																
0	**none**																						

Figure 3. Distribution and comparison of criteria adopted at Frick Art Reference Library

can be found. In particular, the usage of bibliography is consistent with the original rating, especially when compared to scholars' attributions (accepted 298 times and discarded only 18 times). The criterion "archival classification" is consistently used when compared to lower rated criteria, while it is less consistent when compared to bibliography (accepted 41 times and discarded 13 times). This scenario confirms the predominant role of the subjective decision taken by cataloguers during the cataloguing process.

Photo archives comparison. The distribution of criteria chosen by data providers for supporting accepted attributions (regardless competing attributions are recorded) provides insights on art historical data providers' policies. We analyse all records in the three datasets so as to understand whether the rating itself is quantitatively consistent with domain experts' opinion. Figure 4 illustrates the distribution of criteria adopted by the three photo archives (in percentage).

The scenario confirms some results of prior comparative studies. The criterion archival classification is the most used in all of the three archives (30%, 39%, 49%), along with bibliography, which appears to be the main source of information in most of the cases (49%, 8%, 43%). Despite

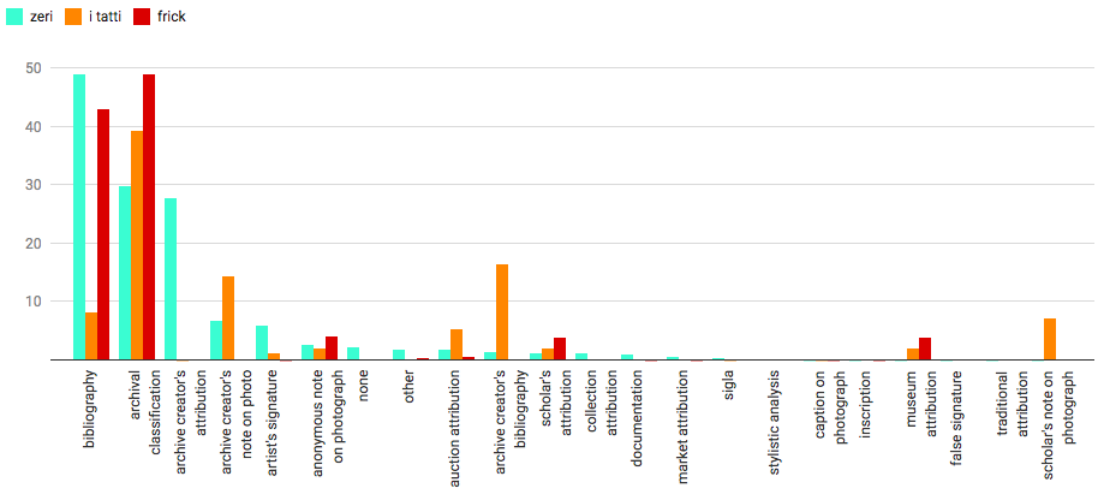

Figure 4. Distribution of criteria in Zeri, I Tatti, and Frick photo archives

the latter could be deemed authoritative criterion *a priori*, an accurate analysis on the validity of such references over time and the cognitive authority of cited authors would deserve more attention (see S6). The third aspect characterizing providers is the significant number of citations gathered by cataloguers, including official statements or notes recorded on the back of photographs. Some providers may prefer certain types of sources rather than others. For instance, museum attributions are well represented in I Tatti (2%) and Frick (4%), while are underrepresented in Zeri (0%). In turn, Zeri relies on market attributions (1%), which seem to be absent in the other two datasets. Auction attributions are mainly cited by I Tatti (5%) and Zeri (2%), and less in Frick (1%).

In conclusion, we revise the original rating and we include the three new criteria emerged from the data analysis, namely archive creator's bibliography, archive creator's attribution and archive creator's note on photograph. These are included in the rating according to archivists' preferences shown in data. We assume the proposed rating of twenty two criteria is valid over the three photo archives and we propose it as a minimum common denominator when comparing contradictory attributions. In Table 1 is listed the final controlled vocabulary of criteria and related rating. The final rating is based on the original domain experts' rating normalized between 1 and 10 to balance its importance among the set of dimensions described in the next section.

Definition of IQ Measures for Addressing Textual and Cognitive Authoritativeness

The corpus analysis shows that a rating of criteria is not always consistent, nor sufficient to address the most authoritative attribution, since other factors may affect reliability of statements. In the second phase of the study we investigate a broader set of dimensions and metrics with regard to textual and cognitive authoritativeness. In particular, we (a) survey a number of IQ measures that apply to the Arts field for measuring textual authoritativeness, and (b) design bespoke metrics for addressing cognitive authority of scholars cited as primary sources of attribution.

S5. Survey of IQ dimensions and metrics. We select a number of existing domain-dependent and domain-independent measures and metrics from (Naumann and Rolker 2000). Dimensions are pruned so as to include only measures that apply to the Arts domain. The selection is made according to online guidelines (Baca & Harpring, 2006), domain experts' consultancy, and aspects highlighted by the corpus analysis. For each dimension we define an assessment method. In

N.	Term	Score
1	Documentation	10
2	Artist's signature	10
3	Archive creator's attribution	9
4	Archive creator's bibliography	8
5	Bibliography	7
6	Archival classification	7
7	Archive creator's note on photograph	7
8	Scholar's attribution	6
9	Museum attribution	5
10	Scholar's note on photograph	5
11	Inscription	5
12	Sigla	5
13	Auction attribution	4
14	Collection attribution	4
15	Market attribution	4
16	Traditional attribution	4
17	Stylistic analysis	3
18	Anonymous note on photograph	3
19	False signature	2
20	Caption on photograph	2
21	Other	2
22	None	1

Table 1
The controlled vocabulary of criteria and the rating

detail, we selected two Subject criteria (features that depends on the observer's perspective), namely relevance and reputation, and two Object criteria (features that characterise the information source), that is, reliability and timeliness.

Relevance is the extent to which information is applicable and helpful for the task at hand. We rely on a list of data providers which are likely to include the attributions. In the proposed use case, the list of data providers includes the three aforementioned art historical photo archives and three multipurpose datasets, namely: DBpedia, Wikidata and VIAF. A common belief in the Arts field is that the more sources agree on a certain attribution, the more such an attribution is likely to be the most relevant among the contradictory ones. We measure relevance by counting the number of sources in agreement on a certain attribution.

Reputation is the extent to which information is highly regarded in terms of source or content. We assume that reputation of information can be inherited by data providers' reputation. Data providers' reputation can be evaluated by relying on third party opinions. In particular, providers that are part of the aforementioned list of data providers are flagged as domain experts or non-experts. Secondly, reputation of cited sources, i.e. historians, is measured by means of two bespoke metrics for measuring cognitive authority (see S6).

Reliability is the extent to which information is correct and trusty. According to domain experts, reliability of criteria motivating an attribution are the most important means to validate

its reliability. We measure reliability of an attribution by using the rating of twenty two criteria extracted from the three surveyed photo archives catalogues.

Timeliness is the distance between the date of the information and the retrieval date. Another common belief in the Arts domain is that the latest recorded attribution - assuming it is also well-documented - is likely to be the most reliable. Timeliness of an attribution is measured by calculating the difference between the date of retrieval and the date of the attribution itself.

S6. Metrics for cognitive authority. When attributions are supported by scholars' attributions, their reputation must be assessed too. Citation indexes for representing scholars' authoritativeness are selected and tuned so as to measure the likelihood of art historians to be reliable sources of information. In particular, the *artist-related index* and the *acceptance-rating* of the scholar are developed.

The *artist-related index* is inspired by the h-index metric. H-index is a metric that uses the number of an author's publications along with the number of times those publications have been cited by other authors in an attempt to gauge an author's perceived academic authority in their given fields of research (Mitchell et al., 2011). The h-index of most of art historians is not available, since they belong to the first half of the 20th century. Moreover, scholars are acknowledged in many ways other than their bibliographic references, such as "verbal communication" or "note on the photograph". In order to apply a citation-based metric to art historians, the following two parameters are taken into account:

- The number of artists to whom the scholar ascribed some artworks. The number is currently limited to artists retrieved in the three photo archives, whose artworks where ascribed by the scholar at least once (discarded attributions that cite the scholar are not counted).
- The number of artworks that the scholar ascribed to a certain artist correspond to the number of the scholar's citations. The number includes all the scholar's accepted attributions retrieved in the three photo archives.

For instance, in the course of his activities Bernard Berenson ascribed artworks to 8 artists. For each of these artists he has been cited as favourite source of attribution respectively 10, 9, 9, 8, 8, 3, 2, 1 times. In details, he has been cited 10 times for having ascribed 10 different artworks to the first artist, 9 times for 9 different artworks to the second artist, and so on. His artist-related index is 5, because he has been cited at least five times with regard to 5 artists. Limits of the metric are evident. Connoisseurs that work on a narrow group of artists, or artists that were not particularly productive, are penalized.

The *acceptance-rating* is a scoped measure that uses the number of a scholar's accepted attributions with regard to a certain artist, along with the total number of possible attributions for that artist (i.e. the total number of artworks surveyed in the three photo archives). Precisely, given a list of tuples (*historian*, *artist*) the rating is calculated for each tuple as the proportion between the number of scholar's citations for that artist over the three photo archives (*numberOfCitations*) and the number of artworks that are ascribed to the latter in the three photo archives (*totalNumberOfArtworks*). For instance, Bernard Berenson has been cited 10 times with regard to Titian's artworks (i.e. 10 of his attributions were accepted by data providers). The three photo archives surveyed 20 Titian's artworks. The acceptance-rating of Bernard Berenson's attributions with regard to Titian is 50%.

IQ Group	IQ Measure	Score	Range
Subject criteria	Relevance	agreement (g)	[between 0 and (n-1)] where n is the total number of retrieved information sources minus the one in scope
	Reputation	domain expert (a)	[0 or 1] boolean
Object criteria	Reliability	criteria (b)	$\sum_{i=1}^{n} x$ where x is the rating associated to a criterion and n is the number of criteria recorded for the attribution at hand
	Timeliness	date (f)	Range: [between 0 and 1]

Table 2
IQ dimensions, scores and ranges

A Model for Ranking Secondary Sources of Attribution

Lastly we apply the developed conceptual framework of IQ measures into a ranking model. As aforementioned, citation indexes defined in S6 do not affect the ranking of results, while these are served along with ranked results offering insights to users.

S7. Ranking model. The Ranking model elaborates a number of steps and incrementally associates a score to attributions recorded in data sources (both accepted and discarded). Different units of measure apply to the definition of partial scores, hence scores lie on different ranges of values. Table 2 summarizes the four aforementioned dimensions, related scores, and ranges.

Relevance is addressed by the *agreement score* (g), that counts the number of providers in agreement with the attribution at hand minus the selected source. For instance, having six data providers, the range of the agreement score is between 0 (no other sources in agreement) and 5 (all the sources agree with the attribution at hand).

The *domain expert score* (a) is a boolean measure that is 1 when attributed to domain experts and 0 when attributed to non-experts. The score is intentionally low so as to not penalize less scholarly sources, such as DBpedia, Wikidata, and VIAF. Indeed, the latter contribute to highlight the broad acceptance of an attribution.

Reliability is measured by relying on the rating of criteria that motivate the attribution. According to domain experts' opinions, the *criteria score* (b) is the one that mostly affects the ranking of results, hence it must weight significantly more than others. The score is cumulative, meaning the sum of all ratings of each criterion supporting the attribution at hand.

Finally, timeliness is measured by the *date score* (f), obtained by comparing the dates of retrieved attributions. The score is normalized between 1 and 0 so as to balance the rating of criteria with a lower rating, e.g. the most recent scholar's attribution should weight as much as the archival classification.

User Study: Measuring Textual and Cognitive Authority of Authorship Attributions

The soundness of the conceptual framework of IQ measures and the ranking model are validated by means of a user study. A proof-of-concept web application called *mAuth - Mining Authoritativeness in Art History*[9], is developed to perform the survey. The application allows users to input the URL of a cataloguing record describing an artwork and to browse the sorted list of attributions fetched in the web of data. We designed a task-based evaluation. Users performed three tasks

remotely and filled in an evaluation form[10]. Tasks are designed so as to reproduce three common scenarios in connoisseurship, namely:

1. Gather information on an artwork whose attribution is unanimously accepted. Only one domain expert is found. Three less scholarly sources agree with the same attribution.

2. Gather information on an artwork whose authorship attribution is debated and that is not sufficiently documented. Only two domain experts are found, and both support their choice by citing scholars. However, such scholars have significantly different citation indexes.

3. Gather information on an artwork whose authorship attribution is debated and that is well-documented. Three domain experts are found. Two sources agree on the same artist and provide plenty of documentation. The third source is in disagreement, does not provide evidences, and is the oldest attribution.

For all of the three scenarios we measured a number of parameters. For the sake of brevity we discuss here only three measures for assessing the User Satisfaction, namely: the User Satisfaction (US) measure, the Rank Satisfaction Score (RSS), and the Perception of Authoritativeness Score (PAS). The US measure measures whether the proposed retrieval process is useful and sufficient to assess the goodness of an authorship attribution. Users were asked to answer the question "Was it easy to find sufficient information for validating the most authoritative authorship attribution?". The RSS score measures user's satisfaction with respect to the order of results and the score associated to each information source. To evaluate the RSS measure, users were asked to answer the question "Do you agree with the ranking of results (i.e. the score attributed to each provided attribution and the order in the list)?". The PAS measure is based on the Net Promoter Score (Reichheld & Markey, 2011) that measures whether a user would prefer and suggest the most rated attribution as the most authoritative one. To evaluate the PAS measure, users answered the question "Do you agree with the suggested attribution?". Participants provided the US, the RSS and the PAS measure by using a Likert scale from 1 to 5 (Strongly disagree to Strongly agree). For all of the three measures we calculated the inter-raters agreement by means of the Fleiss Kappa measure (Fleiss, 1971). Lastly, we collected users' feedback for improving the ranking model. Users were asked to select one or more dimensions that in their opinion would affect the ranking.

We collected feedback from 31 users. The background of participants is the key element of the evaluation. Users mainly belong to some of the most important cultural institutions dealing with art historical data. Other stakeholders in the Humanities and Computer Science were involved to get feedback from different points of view. Domain experts are expected to evaluate the goodness of ranked attributions, while non-domain experts are expected to provide feedback on the soundness of the conceptual framework as applied to any kind of pieces of information, and show whether there are similarities between the art historical research approach and other fields. Table 3 shows users grouped by background and affiliation.

Results and Discussion

Results of the survey are available in (Daquino, 2019b). In Figure 5 are illustrated the US, RSS, and PAS measures for each scenario.

As expected, the US is high in the first and third scenario (84% of user either agree or strongly agree), since the first artwork is unanimously ascribed to the same artist, and the third presents plenty of evidences supporting an attribution rather than others. In the second scenario the US is significantly lower (58%) since attributions are less documented, there are only two sources, both are supported by scholars' opinions, and there is no agreement.

Background	N.	Affiliation
Art historian	1	Warburg Institute
	1	Max Planck Inst. for Art History
	1	Frick Art Reference Library
	1	University of Padua
	1	University of Bologna
	4	Italian Public Education System
	1	Getty Research Institute
	1	University of Rome
Collection manager	1	Getty Research Institute
	1	Yale Center for British Art
	1	Italian Ministry of Cultural Heritage and Activities (MiBACT)
	1	Not specified
	1	Paul Mellon Centre for Studies in British Art
Photo archivist	1	Federico Zeri Foundation
	1	Kunsthistorisches Institut in Florenz
	1	Bibliotheca Hertziana - Max-Planc Institut
	1	Italian Ministry of Cultural Heritage and Activities (MiBACT)
	1	University of Trieste
DH scholar	1	University of Bologna
	1	University of Lausanne
Computer Scientist	2	University of Bologna
	1	Vrije Universiteit Amsterdam
	1	Knowledge Media Institute - Open University
Other	2	University of Milan
	1	University of Florence
	1	University of Bologna

Table 3
Population of the User study

When evaluating RSS, we see that in the first scenario 74% participants either agree or strongly agree; in the second scenario only 38,7% either agree or strongly agree, while 35,5% neither agree nor disagree, and 25,8% disagree; in the third scenario 81% either agree or strongly agree.

When evaluating PAS, in the first scenario we see that 84% either agree or strongly agree; only 42% either agree or strongly agree in the second scenario, while 51,6% neither agree nor disagree; 71% either agree or strongly agree in the third scenario.

The kappa measure is calculated for the 31 raters that evaluated the three cases according to the five categories of the Likert scale: kappa is 33% when evaluating the US measure, 34% for the RSS measure, and 36% for the PAS measure, indicating a fair agreement between raters.

Results show that textual authoritativeness is sufficient in few common scenarios, namely: (1) when there is an agreement between all the sources (first scenario), (2) when there is a disagreement but one source is more documented than others (third scenario), (3) when citation indexes confirm the rating based on textual authoritativeness (third scenario). Limits of our approach are highlighted

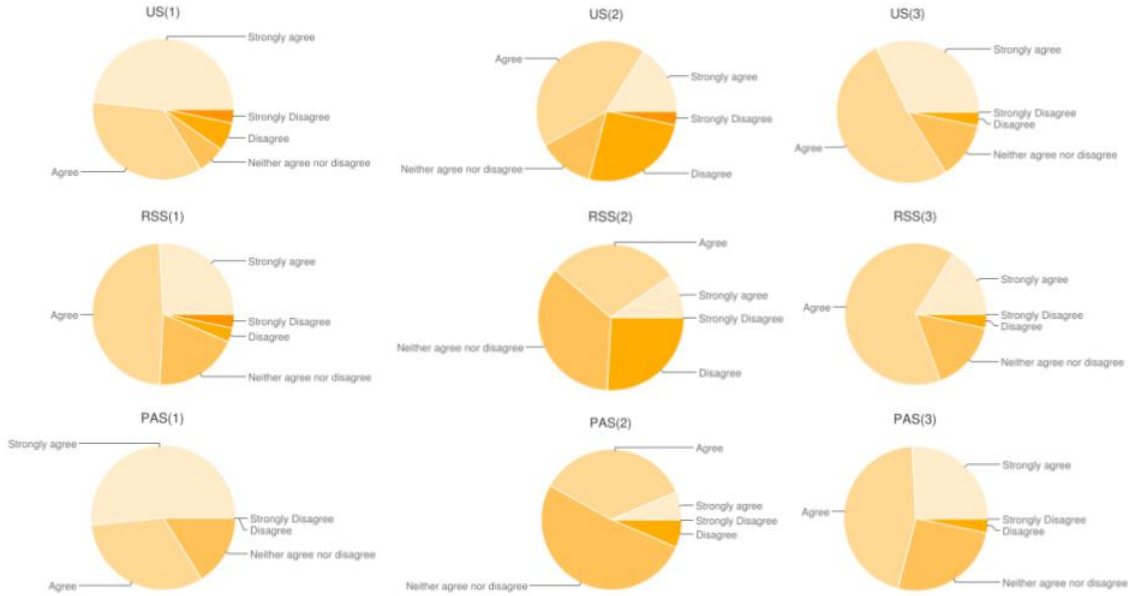

Figure 5. US, RSS, and PAS measures in three evaluated scenarios

by results in the second scenario, namely: (a) when sources are not well-documented, (b) all the sources rely on scholars' opinions, and (c) citation indexes contradict the rating based on textual authoritativeness.

In order to corroborate the assumption that cognitive authoritativeness is the key element when textual authoritativeness is not sufficient, we collected users' feedback on the dimensions they deem relevant. At the end of each task participants were asked to answer the following question "Which criteria would you deem relevant to rank results?". In all of the three scenarios the most voted dimension is scholars' cognitive authority, i.e., "the cited scholar or source of information is considered authoritative (i.e. s/he an high h index)" (74,2%, 61,3%, and 64,5%), followed by the rating of criteria "the source or the criteria underpinning the attribution are the most reliable" (58,1%, 54,8%, and 61,3%), and the data provider's reputation, i.e. "the attribution is provided by a domain expert" (67,7%, 51,6%, and 58,1%).

In summary, users' perception on the current ranking model and the conceptual framework is positive when cited scholars' cognitive authoritativeness is not fundamental for the sake of the judgment, or when it confirms the ranking. Improvements in the ranking model will have to take into account situations where textual authoritativeness is not sufficient. However, providing reliable and comprehensive citation indexes in the Arts and Humanities is challenging and will deserve attention in future works.

Conclusion and Future Work

In this work we presented the research design, methods, and results of a computational analysis performed on art historical linked data. The objective is to assess authoritativeness of secondary sources recording artwork attributions. Results demonstrate that combining domain experts' consultancy and data analysis is sufficient to develop a conceptual framework of IQ measures able to

assess textual authoritativeness of contradictory statements. However, textual authoritativeness is not sufficient when contradictory sources rely on scholars' authoritativeness only. Currently, aspects characterizing cognitive authority are hard to be addressed due to the lack of citation indexes and bespoke measures for assessing art historians' authority. In future works we aim at filling this gap, by collecting and analyzing significant amounts of bibliographic data in the field of Arts so as to explore historical citation networks, develop metrics for measuring scholars' authoritativeness, and include the latter in the ranking model appropriately. Secondly, we aim at involving new data providers, so as to eventually refine the rating of criteria and balancing the ranking model iteratively. In particular, cultural institutions that do not include motivations in their data but that are cited as sources of attribution, e.g. museums, are currently penalized by such a ranking model. We will analyse and compare sources that do not provide detailed information on attributions so as to understand how these influence the art historical debate, and we will tune scores for domain experts.

References

Baca, M., & Gill, M. (2015). Encoding multilingual knowledge systems in the digital age: the getty vocabularies. *Knowledge Organization*, *42*(4), 232–243.

Baca, M., & Harpring, P. (2006). *Categories for the description of works of art*. J. Paul Getty Trust.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 16.

Brilliant, R. (1988). How an art historian connects art objects and information.

Coburn, E., Light, R., McKenna, G., Stein, R., & Vitzthum, A. (2010). Lido-lightweight information describing objects version 1.0. *ICOM International Committee of Museums*.

Cooke, A. (1999). *Authoritative guide to evaluating information on the internet. neal-schuman netguide series.* ERIC.

Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J. J., & McBride, B. (2014). Rdf 1.1 concepts and abstract syntax. *W3C recommendation*.

Daquino, M. (2019a). *mauth - corpus analysis.* (data retrieved from Figshare, `https://doi.org/10.6084/m9.figshare.7411262`)

Daquino, M. (2019b). *mauth - results of the user study.* (data retrieved from Figshare, `https://doi.org/10.6084/m9.figshare.7409384`)

Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., & Vitali, F. (2016). Zeri photo archive rdf dataset.

Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., & Vitali, F. (2017). Enhancing semantic expressivity in the cultural heritage domain: exposing the zeri photo archive as linked open data. *Journal on Computing and Cultural Heritage (JOCCH)*, *10*(4), 21.

Doerr, M. (2003). The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, *24*(3), 75.

Doerr, M. (2009). Ontologies for cultural heritage. In *Handbook on ontologies* (pp. 463–486). Springer.

Farahat, A. O., Chen, F. R., Mathis, C. R., & Nunberg, G. D. (2007, March 6). *Systems and methods for authoritativeness grading, estimation and sorting of documents in large heterogeneous document collections.* Google Patents. (US Patent 7,188,117)

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Freedberg, D. (2006). Why connoisseurship matters. *Munuscula Amicorum: Contributions on Rubens and his colleagues in honour of Hans Vlieghe*, *1*, 29–43.

Ginzburg, C. (1979). Roots of a scientific paradigm. *Theory and Society*, *7*(3), 273–88.

Hevner, A., March, S., Park, J., & Ram, S. (2004). *Design science in information systems research. mis q 28 (1): 75–105.*

Knight, S.-a., & Burn, J. (2005). Developing a framework for assessing information quality on the world wide web. *Informing Science*, *8*.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). Aimq: a methodology for information quality assessment. *Information & management*, *40*(2), 133–146.

McKenna, G., & Patsatzi, E. (2007). *Spectrum: The uk museum documentation standard*. Museum Documentation Association.

Mitchell, G. R., Church, S., Bartosh, T., Godana, G. D., Stohr, R., Jones, S., & Knowlton, A. (2011). Measuring scholarly metrics.

Moreau, L., Groth, P., Cheney, J., Lebo, T., & Miles, S. (2015). The rationale of prov. *Web Semantics: Science, Services and Agents on the World Wide Web*, *35*, 235–257.

Morelli, G., & Richter, L. M. S. (1883). *Italian masters in german galleries*. G. Bell and Sons.

Moro, L., Mancinelli, M., & Negri, A. (2017). Il ruolo dell'iccd nella diffusione dei modelli descrittivi del patrimonio archeologico. *established by: Mauro Cristofani and Riccardo Francovich*(Supplemento 9), 35–46.

Naumann, F., & Rolker, C. (2005). *Assessment methods for information quality criteria*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik.

Park, J.-R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & classification quarterly*, *47*(3-4), 213–228.

Peroni, S., & Shotton, D. (2018). The spar ontologies. In *International semantic web conference* (pp. 119–136).

Reichheld, F. F., & Markey, R. (2011). *The ultimate question 2.0: How net promoter companies thrive in a customer-driven world*. Harvard Business Press.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American society for information science and technology*, *53*(2), 145–161.

Walton, D. (2013). *Argumentation schemes for presumptive reasoning*. Routledge.

Wilson, P. (1983). Second-hand knowledge: An inquiry into cognitive authority.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, *7*(1), 63–93.

Footnotes

[1] https://www.europeana.eu

[2] http://pharosartresearch.org

[3] http://www.fondazionezeri.unibo.it/en

[4] https://itatti.harvard.edu/berenson-library

[5] https://www.frick.org/research/library

[6] (Cyganiak et al., 2014)

[7] http://purl.org/emmedi/mauth

[8] The code realised to perform the analysis and a dump of the datasets are available in `https://github.com/marilenadaquino/mauth/tree/master/data`. A long-term preservation dump of the Zeri dataset is stored in (Daquino, Mambelli, Peroni, Tomasi, & Vitali, 2016). Results of the analysis are available in (Daquino, 2019a).

[9] http://purl.org/emmedi/mauth/search

[10] https://goo.gl/forms/xDLwvCCaEFWm4D5h2