On the Choice of Vocabularies for Archival Data

Veruska Zamborlini[1], Leon van Wissen[2], Albert Meroño-Peñuela[3] and Charles van den Heuvel[4]

[1]*Ontology & Conceptual Modeling Research Group (NEMO), Federal University of Espírito Santo (UFES), ES, Brazil*
[2]*University of Amsterdam (UvA), Amsterdam, Netherlands*
[3]*Informatics Department, King's College London, London, United Kingdom*
[4]*Knowledge & Art Practices, Huygens Institute, Amsterdam, Netherlands*

Abstract

Time and time again researchers are faced with the issue of choosing the most appropriate vocabulary for publishing archival data, particularly in the Semantic Web. Options range from most popular ones, such as schema.org, or more comprehensive ones such as CIDOC-CRM. There are pros and cons in each of them, but no guidelines on how to decide about it. This paper aims at providing some guidance based on an analysis of data at hand but also the requirements of data providers and users. For example, archives often refrain to add much interpretation by providing simple access to categorised documents with simple annotations such as person's names or location names. Moreover, the archival data as well as its digitized versions may present subtleties, such as is the document original or has it been modified, simplified, copied or translated, which is often omitted. Therefore, depending on how much detailed information is actually accessible, but also what are the requirements of the data providers/users, the data can be "placed" at different levels of content literacy/granularity and provenance. By having a clear understanding of the possibilities and limitations of each level, the choice of one or more vocabularies are down to the one(s) that should provide the necessary expressiveness. Naturally, choosing more than one vocabulary also requires some integration task.

Keywords

Archival Data, Vocabulary reuse, Provenance

1. Introduction

When publishing data in the Semantic Web, researchers are often faced with the challenge of choosing the most appropriate vocabulary. This is far from a simple task for a few reasons: (i) it is not easy, and even not recommendable, to build it from scratch; (ii) there are a variety of options to choose from and (iii) it profoundly impacts the expressiveness, the interpretation and the connectivity of your data on the web.

This problem is not exclusive for archival data, but this paper focuses specifically on this kind of data that come from historical documents, records, and artifacts preserved in various national and regional archives. Researchers often struggle with dilemmas: choosing for the most popular vocabulary, such as schema.org, or more comprehensive ones, such as the CIDOC-CRM, yet

YODA'24: ContemporarY Ontologies for Digital Archives Workshop, July 15–19, 2024, Enschede, Netherlands
veruska.zamborlini@ufes.br (V. Zamborlini); l.vanwissen@uva.nl (L. v. Wissen); albert.merono@kcl.ac.uk (A. Meroño-Peñuela); charles.van.den.heuvel@huygens.knaw.nl (C. v. d. Heuvel)
0000-0003-3425-0481 (V. Zamborlini); 0000-0001-8672-025X (L. v. Wissen); 0000-0003-4646-5842 (A. Meroño-Peñuela); 0000-0001-9638-400X (C. v. d. Heuvel)

something in between. An extensive list of vocabularies can be found on the report of a study by the European Commission for management, exchange and publication of archival data [1]. This study highlights that (i) providing data using the Semantic Web framework has the potential to make data more accessible and interoperable, and that (ii) although a consensus on a body of standard vocabularies to use exists, a heterogeneity of practices needs to be acknowledged. In fact, there are pros and cons in each vocabulary, but no guidelines on how to decide which ones to choose.

This paper aims at providing some guidance for this choice by looking at data in two dimensions of content granularity/literacy and provenance. By having a clear understanding of the possibilities and limitations of each level in those dimensions, the choice of one or more vocabularies is down to the one(s) that should provide the necessary expressiveness, also taking into account the conventions and most commonly used vocabularies in the field. An analysis to position the data at hand in those dimensions not only reckons with features of data but also the requirements of both data providers and users.

As an example for requirements of data providers, archivists often provide only access to categorised documents with simple annotations such as person's names or locations (called indexes), leaving out deeper and possibly arguable interpretations of the archival resources. After all, originally these indexes on the material were made for easier access to the material for researchers and genealogists that would further read and investigate the contents of the respective material anyway. Time-wise this was possible, because only a few records had to be inspected. As an example for requirements of data users, nowadays, researchers and data scientists are interested in analysing thousands of archival documents at the same time, requiring other means of modelling the data and also of obtaining it (e.g. through Handwritten Text Recognition and Natural Language Processing).

Common features of historical data in general are uncertainties and incompleteness. Interpretation is mostly unavoidable and users may require detailed description of it. For example, information about the origin and the manipulations of the archival resources in multiple versions, such as copies, simplifications, translations etc. are often omitted, because it is not (fully) known or because the providers lack the means to express what is known. However, those processes may add several layers of interpretation that ideally should be made accessible.

Naturally, it may be necessary to consider using more than one vocabulary, which would bring its own challenges. Semantic interoperability is far from obvious, and might require a proper ontological analysis of the choices (often implicitly) made in each vocabulary to avoid misleading integration. For example, in [2] the authors show that concepts such as period and temporal entity do not have the same meaning in CIDOC-CRM, OWL-Time and PeriodO vocabularies[1] through an ontological analysis. Moreover, vocabularies may choose different modelling styles that are also not trivial to integrate. For example, event-based CIDOC-CRM models birth (CRM:E67_BIRTH) as a class while attribute-based schema.org models birth (SCHEMA:BIRTHDATE) as a property that expects a literal value. Mistaken alignments of classes and/or properties lead to logically invalid models or to misleading conclusions, that even worse are not detectable by a reasoner.

[1]Respectively accessible at www.cidoc-crm.org , www.w3.org/TR/owl-time and https://perio.do/en/

The strategy[2] here proposed was developed in the context of the Golden Agents project[3] aimed at analyzing interactions between the production and consumption of cultural goods (e.g. paintings, books, silverware) during the long Dutch Golden Age (ca. 1570s-1750) using linked data from various cultural heritage institutions. To understand this consumption and explain the flourishing of Amsterdam's cultural industries, millions of digitized archival documents and metadata records from the Amsterdam City Archives (SAA) with data on producers, consumers and the goods that were interchanged were brought together as linked data.

In the remainder of this paper, Section 2 presents the related work, Section 3 presents a case study motivating the data analysis strategy proposed in Section 4, and Section 5 presents discussion and future work.

2. Related Work

Almost every knowledge engineering methodology identifies the reuse of existing ontologies and vocabularies as a crucial step when modelling and publishing Linked Data on the Web [3, 4, 5, 6, 7]. However, there is a broad landscape in how exactly this reuse must be enacted [8]: for example, some methodologies recommend a direct or indirect reuse [9], or based on design patterns [10, 11], etc. More recently, the FAIR data principles [12] have established a framework for making data Findable, Accessible, Interoperable and Reusable; with nanopublications [13, 14] being a successful implementation that distinguishes between an *assertion* (i.e. observations, literal content), *provenance* (i.e. backing references, workflows; typically modelled using the W3C PROV vocabulary [15]) and *publication info* (i.e. metadata, documents/collections). Although this addresses some of our challenges, nanopublications (a) have mainly been used in the life sciences domain, where there is often little space for interpretation compared to archival data; (b) do not address data reconstruction explicitly, as often relate to the result of contemporary scientific processes; and (c) do not establish different degrees or levels of required provenance. Similarly to nanopublications, the Dutch Historical censuses (CEDAR) [16] deploy a similar approach for historical data that distinguishes between *raw* observations, *annotations* that are subject to interpretation, and *statistical* data cubes (using the RDF Data Cube vocabulary [17]). However, this focuses on the specific domain of historical statistics, rather than offering an abstract framework for any archival data; and records only provenance of data transformation processes, rather than historical processes preceding archival documents. Various vocabularies exist that are specifically designed to model cultural heritage objects (e.g. CIDOC CRM [18]), bibliographic information (e.g. FRBR [19]), and glossaries for archival records [20]. Unfortunately, these are all specific to their corresponding domains, and do not extend to more general notions of provenance.

[2] A first version was discussed at LODLAM Summit 2020 https://lodlam.net/challenge-entries/ and also later at RSA 2022 https://rsa.confex.com/rsa/2022/meetingapp.cgi/Paper/13201).

[3] Golden Agents: Creative industries and the making of the Dutch Golden Age www.goldenagents.org funded by Dutch Research Council (NWO) www.nwo.nl.

3. Case Study

An archival organization could provide digitized information about archival records as follows:

1. indexed information, i.e. just key information and metadata of the record (not the full text), e.g. a notice of marriage between John and Mary on a certain date;
2. scanned books/documents under some classification, e.g. all the marriage banns at the city hall;
3. index information connected to the scanned page in which it appears;
4. full text extraction from a record.

Let's consider as a case study a record documenting the notice of marriage between Jan Ceuler and Susanna de Bock, taken from the Amsterdam City Archives (SAA) and developed during the Golden Agents project. Table 1 illustrates raw data available as items 1-4 for the mentioned example. It shows (1) indexed data including name, location, date and others, it also shows (2) a scanned page containing the referred record, (3) the URL connection between the index data and the scanned page and (4) the full text extracted from the record providing the name, age and location of each person with the intention to marry the other and of their witness.

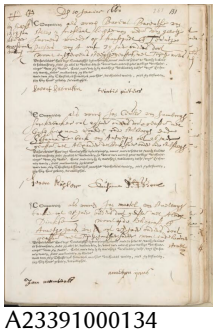
(1)

Index type:	Notice of Marriage
Authority:	*De Pui*
Book:	*684*
Page:	*261*
Location:	*Amsterdam*
Date:	*10/01/1660*
Groom:	*Jan Ceuler*
Bride:	*Susanna de Bock*

(2)

Index x Scan:	*https://archief.amsterdam/...*

(3)



A23391000134

(4)

Compareerden als vooren Jan Ceuler van Hamborgh Suijckerbacker out 29 jaar ouders doot geassisteert met Gose Sirixma wonende inde Stilsteegh ende Susanna de Bock van Antwerpen out 26 jaar geassisteert met Alexander de Bock haer vader inde Stilsteegh Joan Köhler Susanna de Bock

Table 1
Example of raw data illustrating (1) indexed data, (2) a scanned page, (3) a URL connecting a record's index data and the scanned page where it could be found https://archief.amsterdam/indexen/deeds/4aa6a4e1-cd13-4586-8993-af9fe3523fd6?person=961f6b17-36d8-53f7-e053-b784100aa83b and (4) full text extraction from the same referred record.

Observe that users are not provided with information on its provenance and reliability. How and by whom were the indexes and scans created[4] and published and who and in which ways extracted the texts. And more importantly for provenance (and helpful for end-users as a visual aid), where in the scan can information be found? In order to provide more reliable

[4]Due to legacy at the archives, this would be impossible for old indexes. However, nowadays it is very much needed (e.g. due to any bias that can be present in data) to report on the coming to being of indexes. What is more, it is good to know whether the data came from the archive itself, any external research project, or by crowdsourcing and citizen science. Both for attribution as well as to understand the modelling choices or classifications that were made (and from what perspective).

data, ideally the users would have access to (i) the source of the data and (ii) the production processes of the data. Both are often referred to as provenance data, in addition to what is often called provenance in cultural heritage, referring to the processes of creation and ownership of the cultural object (see ISO 8000-120: 2016). It should allow for users to inspect and validate the original handwritten data or the original text or whom came with which interpretations. Thereby, this information increases the credibility of the data offered by the archive and the likelihood of being used in research. Some ways in which more provenance data could be provided are:

1. detailed information on the location of an entire record on a scanned page (e.g. by including coordinates of bounding boxes or regions, cf. the IIIF Presentation API and its FragmentSelectors and SVG Selectors);
2. even more detailed information on the location of each indexed information in the record section on the scanned page or the recognized or transcribed text (e.g. by including character offsets);
3. detailed information on the processes of obtaining the scans, the indexed information and/or the full text and their connections, including any changes made to the (metadata) record. .

Figure 1 displays the scan of a double page with the referred notice of marriage record highlighted within a (red) box, at the bottom right. This selection can be produced by indicating the record's xywh pixel coordinates on the scanned page, and be classified as a record section. The chosen vocabulary in this particular example was inspired by Web Annotation vocabulary [21] The raw text of such a section can also be extracted as the content of the record section.

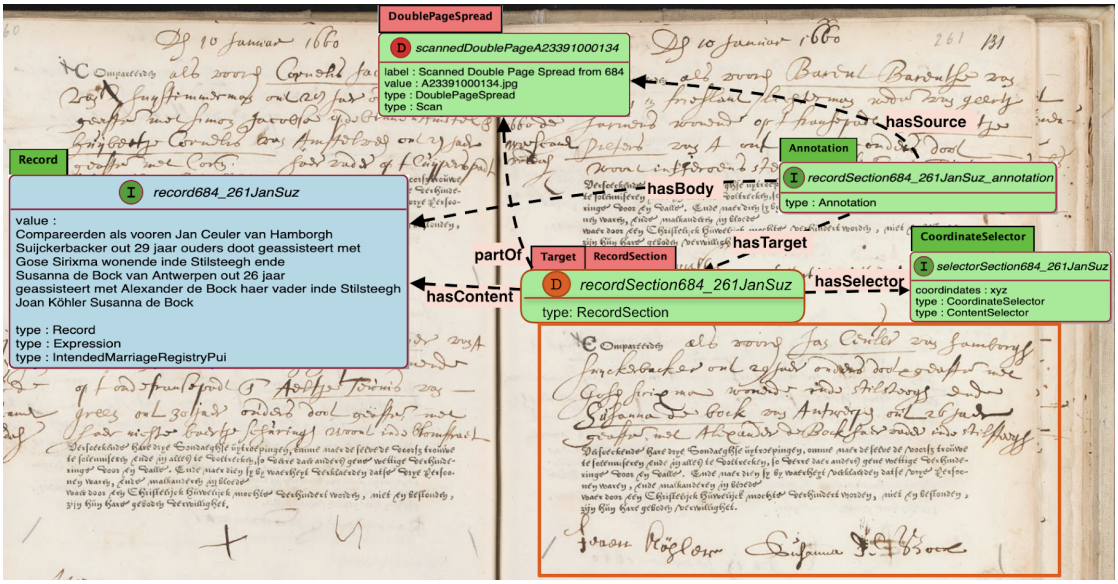


Figure 1: Scanned page showing the Amsterdam marriage banns between Jan Ceuler and Susanna de Bock, with its precise location described as a record section and the raw text extracted as its content.

Moreover, in a similar fashion, more specific sub-sections can be created to indicate where a particular name is mentioned in the record, as depicted in Figure 2, also with its raw content.

As for the information on the archiving processes, Figure 3 depicts the processes of creating and digitizing documents as indexes and scans at the Amsterdam City Archives. The two events depicted more to the left, called coverage events, represent a complex event covering the whole creation of the documents in a collection or book. Naturally, these events, which happened in the 17th century, outputs the original documents and precedes the ones of digitization. First point to highlight is that the current index, provided as open data in XML, was not digitized directly from the documents, but from another (paper or rolodex) index previously created. This is important for understanding that any interpretation, at least regarding the early modern Dutch handwriting, happened in the first process of indexing, not in the second one. Another point is that the process of scanning the archival resources happened separately. Therefore, the connection between the indexed documents and the scanned pages in which they appeared, as well as the extraction and annotation of particular mentions, could only have happened a posteriori in another complimentary digitization process.

Now, from the records and sections one can identify references to (supposedly real world) entities, such as the specific mention of the name Jan Ceuler van Hamborgh. Figure 4 illustrates a detailed extraction of references from the record section, including references within references. Here the vocabulary choices were inspired by Factoid Prosopography Ontology [22], according

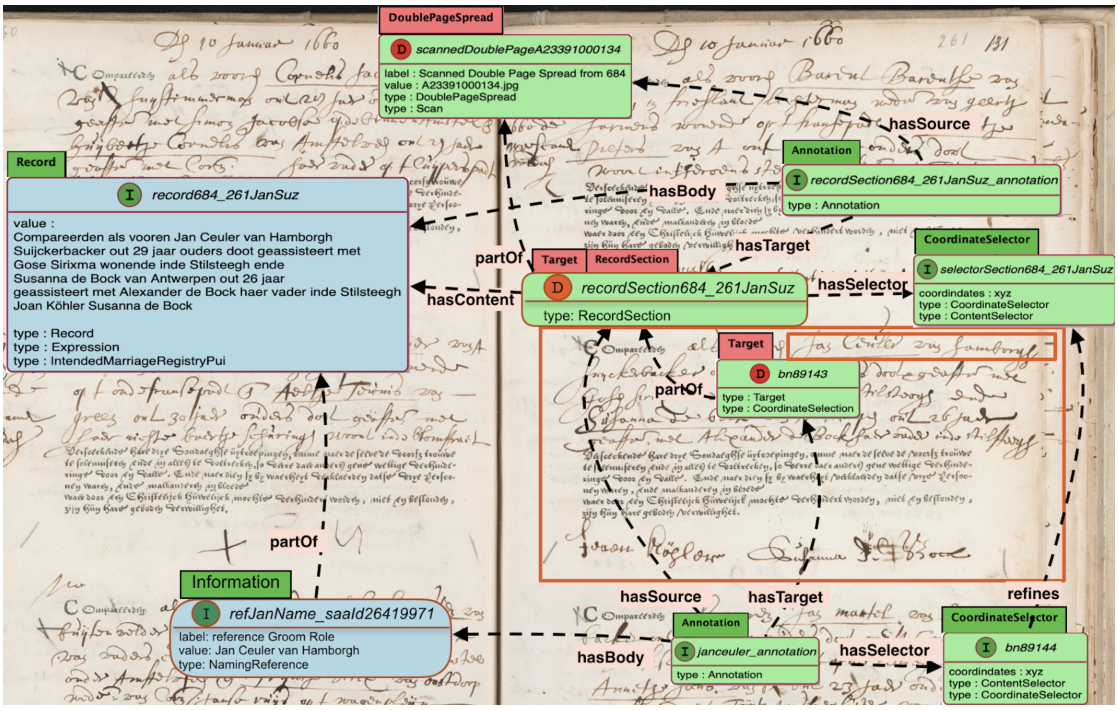


Figure 2: Scanned page showing in particular the mention of a name, Jan Ceuler van Hamborgh, as part of the Amsterdam marriage banns with its precise location described as part of the record section and its content as part of the record content.

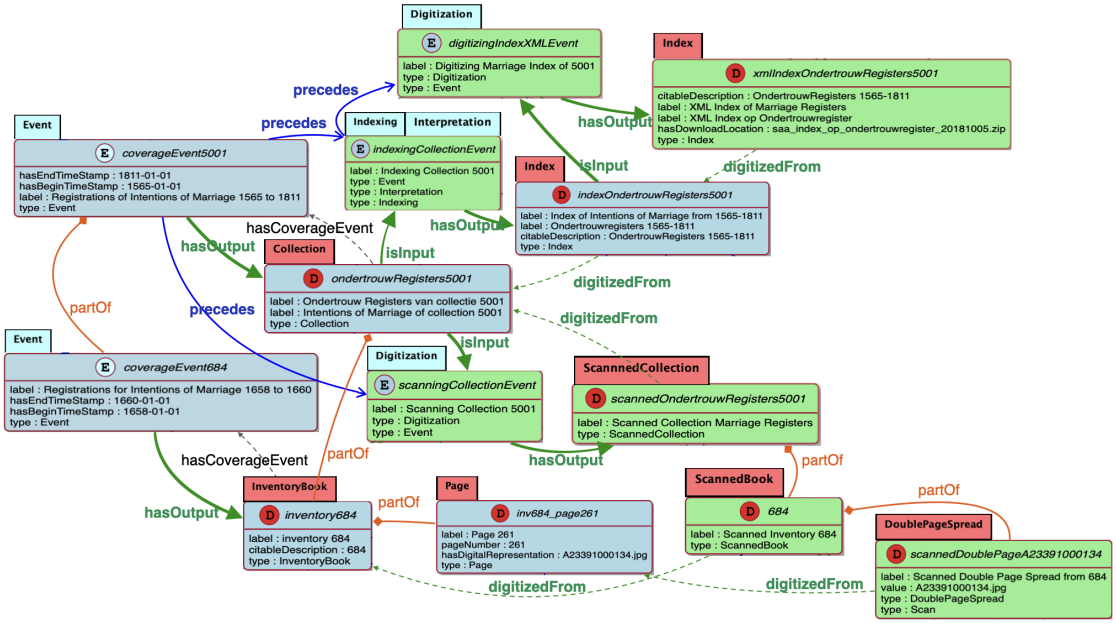


Figure 3: Processes of creation and digitization of documents as indexes and scans from the Amsterdam City Archives.

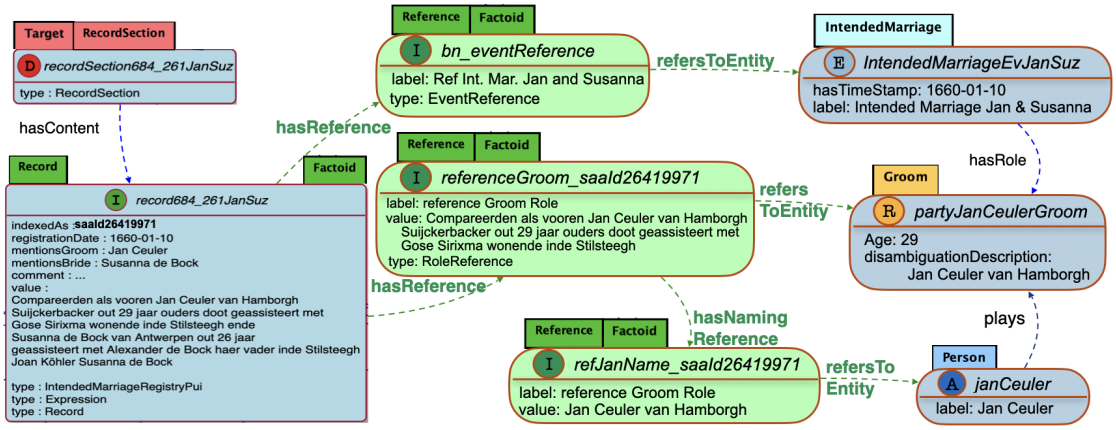


Figure 4: Data can be extracted from a document literally as textual references, and then connected to entities of certain types expecting the mentioned references to obtain in reality, for example, as events in which certain people perform roles like Jan Ceuler as the groom registering his intention to marriage.

to which any piece of information containing references can be called a factoid, including references themselves.

Naturally, one could consider skipping these steps and, for example, connect the record directly to the person-entity named Jan Ceuler. That only means that intermediate steps were made, but left implicit. This simplifies the representation, but also may leave less space to accommodate disagreements. The extracted references are meant to be less arguable, as they

should contain the literal content without any adjustment of language evolution or translations. Meanwhile, those references refer to ”entities” of certain types, sometimes also called observations. At this point, adjustments of the language used are often performed and also reliability of the data is presumed. It is a moment where one could consider that the ”interpretations” would start, which is almost inevitable for historical data. Moreover, different levels of interpretation can take place, more or less ”directly” drawn from the raw-reference data.

Figure 5 shows two interpretations that are not directly stated in the references. The first, assuming a cultural rule, infers that a marriage event might have happened not earlier than 2-3 weeks after the registered notice of marriage. The second assumes that the reference to the groom, Jan Ceuler van Hamborgh, might be related to his city of origin, namely Hamborgh (currently called Hamburg, Germany). It could result in associating Jan Ceuler to a place entity called Hamborgh (back then) as his hometown. This extracted information could potentially provide information on Ceuler's town of origin, even though it was not explicitly stated as his hometown.

Finally, Figure 6 illustrates another level of interpretation, in which entities that are more or less directly drawn from the references, are inferred to be the same. It adds to our running example a baptism record of a child born to a Jan Keulder and Susanna de Bock in April 1668. Using some identity criteria, it is possible to infer that the groom Jan Ceuler is the same as the father Jan Keulder (similarly the mentioned bride and mother could be inferred to be the same). Moreover, it can be concluded [or inferred] that the marriage event that happened few days after the notice of marriage, must be same as the marriage expected to have happened before the baptism of their child.

By combing the information from two archival resources, we are 'reconstructing' the life of

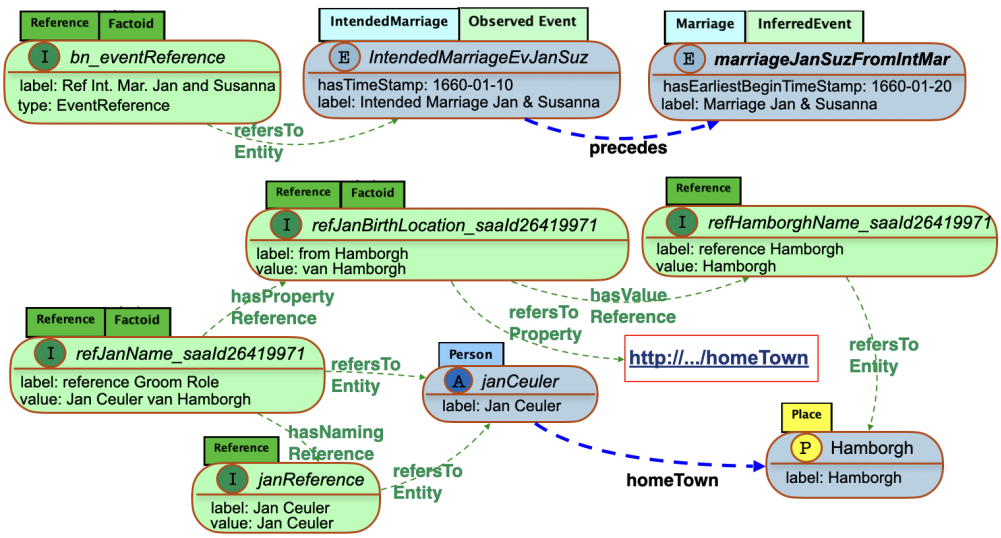


Figure 5: Data can be derived less literally from a document as entities of certain types, mentioned directly or indirectly as references, and expected to obtain in reality, for example, as events such a marriage that is expected to follow an intention of marriage, or a hometown that is mentioned as part of Jan Ceuler's name as begin the city of Hamborgh.

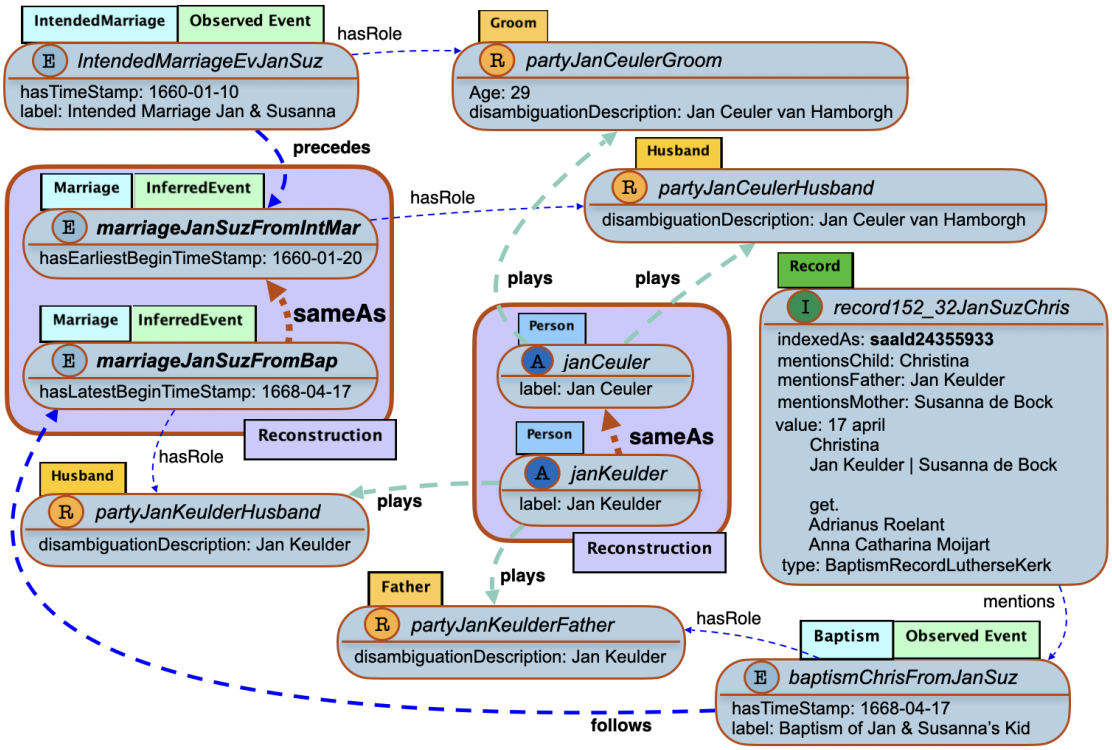


Figure 6: Data from different documents can be combined concluding that they describe events involving the same (reconstructed) entities, such as Jan Ceuler mentioned as groom in the notice of marriage record being the same as Jan Keulder mentioned as father in a baptism record. Moreover, inferred that the marriage event that happened few days after the notice of marriage, must be same as the marriage expected to have happened before the baptism of their child.

these two persons: we are making a reconstruction of the entities based on our observations in individual records. Again, the processes and decision of getting to such conclusions could be documented in detail providing arguments to agree/disagree with. [23].

4. Strategy

An strategy for analysing different "types" of archival data was developed while investigating the issues mentioned in Section 3. It considers two orthogonal dimensions for data details: content literacy/granularity and provenance. Table 2. provides a summary illustrating how the dimensions can be combined in some scenarios.

First, regarding the provenance dimension, three levels are considered: minimum, detailed and advanced. They correspond with having as little as possible to very detailed provenance information. Naturally, there could be more than three.

Minimum provenance means literally as little as possible, or provenance information is not a priority. An example can be seen in Table 1 in Section 3, where indeed nothing is known

about how the data came to be, as well as Figure 6 where nothing is known about how the entities were reconstructed.

Detailed provenance describes the steps and or the processes of obtaining the data. For example, the processes of digitization of a collection, or the specific location of a registry or particular mention on the page of a book or yet the specific sequences of (text) references in which an entity was mentioned. Examples from Section 3 are found in Figures 1 to 5. Regarding integration/reconstructions, this level would require some sort of reification, either for each identity link (same as) or for a group of them (linksets) where more detailed information on the disambiguation process could be described.

	Minimum provenance	Detailed provenance	Advanced provenance
Documents /Collections	Basic information about documents/records, possibly grouped under certain criteria/classifications.	Information that allows access to the original document or yet its scanned version.	Description of the archiving processes of a document/-collection: Who has created the documents? How were them digitized? Are the processes reliable? Are the documents originals or copies/-transcriptions?
Literal Content	Mentions/descriptions of individuals/roles and/or their types.	Annotate where the mentions/descriptions are in the text and/or scan.	Have the mentions/descriptions been modified/adapted/translated in the process? Errors introduced?
Direct Interpretation	Events, roles, objects and properties directly observed in each document.	Break down the content Breakdown of the content into parts and connect to the observed entities and their properties	Are entities, types/roles and properties correctly identified? Are there ambiguities? Could the observation be untrue? Uncertainty level can change.
Indirect Interpretation	Events, roles, objects and properties that can be indirectly inferred.	Why and how extra information can be derived from the ones already extracted or from the original content.	What type of inferencing? How certain is it or what is the probability?
Integration/Reconstruction interpretation	Several observations can be combined to an individual concept through time.	Which specific properties were used for disambiguation? Which process/rules were applied? Has it been validated?	Is the provided information enough for disambiguation? How accurate is the disambiguation process? What is the expertise of the validator?

Table 2
Examples of issues regarding two dimensions of data: content literacy/granularity and provenance.

Advanced provenance considers describing information regarding the reliability of the data, including uncertainties inherent to the data, to the inference rules or yet the ones introduced in the process of obtaining the data. Who created the document? Is the creator trustworthy (e.g. by affiliation or status)? Is it a original document or has it been copied/-translated? There has been any damage/modification to the original document? There can be errors introduced in the process? Is it likely that someone may have lied, for example, about the age, gender or place or origin? Particularly regarding uncertain assertions about individuals, it is still a challenge how this should be represented, either qualitatively or quantitatively. Regarding integration/reconstructions in particular, this would require for identity links (same as) to be reified and qualified with a metric often referred to as similarity.

Second, regarding the content literality/granularity dimension, five levels are considered:

Documents/Collections: includes more general data about documents and their digitization, for example, a book containing baptism registries or yet several books registering baptism events. Here data is not about the content of one document in particular, but regards who, how and when the documents were (re)produced and preserved and their identification within an index. It may indicate in which page(s) of which book a document/registry is, or even its precise position in a scanned page. It also might concern, for example, the church who produced a baptism registration book, which would not indicate the location but also the denomination of the involved people. The location could indicate if the dates are to be interpreted as Gregorian or Julian calendar[5].

Literal content: includes descriptions of the content of document/registries and entities mentioned in it as literally as possible, ideally as references in the text. It may also indicate where exactly the mentions are located in a scanned page. The annotation of such entities' mentions may already indicate the kind of entity mentioned or the role played by them. For example, a document mentions names of people and locations.

Direct interpretation: includes interpretations that are drawn directly from the information/references in the document, regarding events and the roles of the entities. For example, a child baptized on a certain date and the parents.

Indirect interpretation: includes interpretations that are drawn indirectly from the information in the document, also regarding events and the roles of the entities. For example, the name of person could indicate his/her profession and/or his/her place of birth. Or even, a baptism could imply some assumption/approximations about the birth date and location, or burial and death analogously, according to religion and costumes. Traceability regarding the data that were used as evidence for the conclusions, as well as for the processes and uncertainties involved so that more detailed provenance could be provided.

[5]Not all cities in the Netherlands used the same calendar. While Amsterdam used the Gregorian calendar from the end of the sixteenth century onwards, the city of Utrecht for instance kept on using the Julian calendar until the end of the seventeenth century.

Integration/Reconstruction interpretation: includes the identity links connecting mentions from different documents. Traceability regarding the data that was used as evidence for the conclusions, as well as the processes and uncertainties involved are more detailed provenance that could be provided.

It makes clear that a vocabulary choice will have a direct influence on how much details would be possible or necessary and vice-versa.

5. Discussion and Conclusion

This paper presents an strategy meant to support the choice of vocabularies to reuse for digitization of archival data. It proposes understanding the data by looking into two orthogonal dimensions: (i) content literacy/granularity and (ii) provenance. There is not a single answer to the question which vocabulary to use, as it will depend on the level of details in both dimensions in which the data is meant to be described. This implies not only understanding the features of the data at hand but also the requirements of data providers and users. When this is clear, then one or more suitable vocabularies can be chosen as the ones that provide the necessary expressiveness.

The idea was developed during the Golden Agents project where data from the Amsterdam City Archives were analysed in order to be published as RDF data with enrichments. It was influence by the Resource-Observation-Reconstruction structure present in ROAR vocabulary[6] and later further developed as ROAR+[24, 25]. Some interesting things became clear in this process: (1) the city archives' role as data provider was not to provide enrichments that would add much interpretation to the original data, while (2) the project required to enrich data as much as possible, for example integrating mentions to create evidence-based storylines of events for Amsterdamers in the 17th century [23] and ultimately to get insights in the production and consumption of cultural goods of the Dutch Golden Age. Therefore, the data produced under the authority of the SAA would go until a certain level, while the project had to bring it further. Naturally, the vocabulary choices, reflecting different requirements, were not the same.

During the Golden Agents project it was not possible to perform a through analysis of digital humanities vocabularies and how they would correspond with in the proposed dimensions. This would be an interesting future work, although often one vocabulary will not cover completely one level, neither be contained in one. In particular, correspondence of these dimensions with ontologies of provenance and historical assertion records, such as Prov-O[7] and the STAR model developed in the Releven project[8] and specially Records in Context (RIC)[9], a promising ontology for describing archival record resources and their contextual entities and also the reference model from ISO called Open Archival Information Systems (OAIS ISO 14721)[10]. Another important investigation would be an ontological analysis of the general entities present in each level/dimension to support corresponding them with vocabularies. The outcomes of such an

[6] https://www.leonvanwissen.nl/project/roar/
[7] www.w3.org/TR/prov-o/
[8] releven.univie.ac.at/
[9] www.ica.org/resource/records-in-contexts-ontology/
[10] https://www.iso.org/standard/57284.html

analysis implemented would make finding existing ontologies and vocabularies for reuse much easier.

References

[1] E. Comission, Study on Standard-Based Archival Data Management, Exchange and Publication Final Report, Technical Report, Historical Archives Service, 2018. URL: https://ec.europa.eu/isa2/sites/isa/files/isa2_action_2017_01_standard_based_archival_data_management_final_report_v1.00.pdf.

[2] C. van den Heuvel, V. Zamborlini, Modeling and Visualizing Storylines of Historical Interactions. Kubler's Shape of Time and Rembrandt's Night Watch, in: R. P. Smiraglia, A. Scharnhorst (Eds.), Linking Knowledge: Linked Open Data for Knowledge Organization and Visualization, 1 ed., Ergon-Verlag, Baden-Baden, 2021, pp. 99–141. URL: hol. doi:10.5771/9783956506611-99.

[3] K. Kotis, G. A. Vouros, Human-centered Ontology Engineering: The HCOME Methodology, Knowledge and Information Systems 10 (2006) 109–131.

[4] K. I. Kotis, G. A. Vouros, D. Spiliotopoulos, Ontology engineering methodologies for the evolution of living and reused ontologies: Status, trends, findings and recommendations, The Knowledge Engineering Review 35 (2020).

[5] A. d. Moor, P. D. Leenheer, R. Meersman, DOGMA-MESS: A Meaning Evolution Support System for Interorganizational Ontology Engineering, in: International Conference on Conceptual Structures, Springer, 2006, pp. 189–202.

[6] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The NeOn Methodology for Ontology Engineering, in: Ontology engineering in a networked world, Springer, 2012, pp. 9–34.

[7] D. Vrandečić, S. Pinto, C. Tempich, Y. Sure, The DILIGENT Knowledge Processes, Journal of Knowledge Management 9 (2005) 85–96.

[8] V. A. Carriero, M. Daquino, A. Gangemi, A. G. Nuzzolese, S. Peroni, V. Presutti, F. Tomasi, The Landscape of Ontology Reuse Approaches, IOS Press, 2020. URL: http://dx.doi.org/10.3233/SSW200033. doi:10.3233/ssw200033.

[9] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, Arco: The italian cultural heritage knowledge graph, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, Springer, 2019, pp. 36–52.

[10] V. A. Carriero, P. Groth, V. Presutti, Empirical ontology design patterns and shapes from wikidata, Semantic Web (2023).

[11] V. Presutti, E. Daga, A. Gangemi, E. Blomqvist, Extreme design with content ontology design patterns, in: Proc. Workshop on Ontology Patterns, 2009, pp. 83–97.

[12] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[13] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, Information services & use 30 (2010) 51–56.

[14] T. Kuhn, A. Meroño-Peñuela, A. Malic, J. H. Poelen, A. H. Hurlbert, E. C. Ortiz, L. I. Furlong, N. Queralt-Rosinach, C. Chichester, J. M. Banda, et al., Nanopublications: a growing resource of provenance-centric scientific linked data, in: 2018 IEEE 14th International Conference on e-Science (e-Science), IEEE, 2018, pp. 83–92.
[15] P. Missier, K. Belhajjame, J. Cheney, The w3c prov family of specifications for modelling provenance metadata, in: Proceedings of the 16th international conference on extending database technology, 2013, pp. 773–776.
[16] A. Meroño-Peñuela, A. Ashkpour, C. Guéret, S. Schlobach, Cedar: the dutch historical censuses as linked open data, Semantic Web 8 (2017) 297–310.
[17] R. Cyganiak, D. Reynolds, J. Tennison, The rdf data cube vocabulary, W3C recommendation 16 (2014).
[18] M. Doerr, The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata, AI magazine 24 (2003) 75–75.
[19] B. Tillett, What is frbr? a conceptual model for the bibliographic universe, The Australian Library Journal 54 (2005) 24–30.
[20] R. Pearce-Moses, L. A. Baty, A Glossary of Archival and Records Terminology, volume 2013, Society of American Archivists Chicago, IL, 2005.
[21] P. Ciccarese, R. Sanderson, B. Young, Web Annotation Vocabulary, W3C Recommendation, W3C, 2017. URL: https://www.w3.org/TR/2017/REC-annotation-vocab-20170223/.
[22] M. Pasin, J. Bradley, Factoid-based prosopography and computer ontologies: Towards an integrated approach, Digital Scholarship in the Humanities 30 (2015) 86–97. URL: https://www.kcl.ac.uk/factoid-prosopography/overall-concepts.
[23] A. Idrissou, V. Zamborlini, F. Van Harmelen, C. Latronico, Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers, in: K-CAP 2019 - Proceedings of the 10th International Conference on Knowledge Capture, ACM, New York, NY, USA, 2019, pp. 259–262. URL: https://dl.acm.org/doi/10.1145/3360901.3364440. doi:10.1145/3360901.3364440.
[24] L. van Wissen, V. Zamborlini, C. van den Heuvel, Modeling provenance and uncertainties in the use of archival sources of the dutch golden age, 2022. URL: https://rsa.confex.com/rsa/2022/meetingapp.cgi/Paper/13201, presented at the 68th Annual Meeting of the Renaissance Society of America (RSA), 30 March - 2 April 2022, Dublin, Ireland.
[25] L. van Wissen, V. Zamborlini, C. van den Heuvel, Toward an ontology for archival resources. modelling persons, objects and places in the golden agents research infrastructure, 2021. URL: https://dhistory.hypotheses.org/361, data for History Lecture 2021: Modelling Time, Places, Agents, Online.