Græcissāre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin

Greta Franzini, Marco Passarotti,
Francesco Mambrini, Giovanni Moretti
Università Cattolica del Sacro Cuore
CIRCSE Research Centre
Largo Gemelli 1,
20123 Milan, Italy
name.surname@unicatt.it

Federica Zampedri
Università degli Studi di Pavia
Pavia, Italia
federica.zampedri01@universitadipavia.it

Abstract

English. This paper describes the addition of an index of 1,763 Ancient Greek loanwords to the collection of Latin lemmas of the *LiLa: Linking Latin* Knowledge Base of interoperable linguistic resources. This lexical resource increases LiLa's lemma count and tunes its underlying data model to etymological borrowing.

1 Introduction

"Graecia capta ferum victorem cepit"[1]
Horace, *Epistles*, II, 1, 156

Boasting over two thousand years' worth of written attestation, Latin's evolutionary history is among the longest in existence. The diachronic and geographical reach of the Roman Empire exposed Latin, an Indo-European Italic language, to many regional dialects and languages, including Ancient Greek. The mutually profitable linguistic contact between Latin and Ancient Greek[2], facilitated by their similar morphosyntactic structures and characteristic syntheticity (Ledgeway, 2012, pp. 10-28), is most evident in their vocabulary, chiefly calques and loanwords. Both lexemes presuppose a certain knowledge of the donor language, but while the former takes from the donor *with* translation, the latter does not (Hock and Joseph, 2009, p. 252).

Examples of Latin words calqued from Ancient Greek are *unicornuus* "unicorn" (*unus* "one" + *cornu* "horn") from μονόκερως (μόνος "one" + κέρας "horn"), and *infans* "infant" (*in-* "not" + *fans* "speaking") from νήπιος (negative prefix νη- + ἔπος "speech"). Calques can also involve affixes, as is the case of Latin's suffix *-us* being substituted for the Greek *-os* (Hock and Joseph, 2009, p. 253). The adjective "dramatic", for instance, is attested as both *dramaticos* and *dramaticus*.

Example loanwords in Latin are *crocodilus* "crocodile", imported from the Ancient Greek κροκόδειλος, and *liquiritia* "liquorice" from γλυκύρριζα. Adams identifies three categories of Greek loans in Latin (2003, p. 443):

> (1) words for which there existed a Latin equivalent; the writer was so familiar with the local Greek term that he adopted it in response to local conditions; (2) local Greek technical terms for which it might have been difficult to find a Latin equivalent; and (3) transfers determined by a writer's lack of fluency in Latin, as a result of which he either adopted Greek words because he was unaware of their Latin equivalents, or did so unconsciously because of his poor command of Latin.

For each category, Adams provides a handful of examples, including (1) *(h)amaxa* from ἅμαξα "wagon", (2) *buneurum* from βούνευρον "whip of oxhide" and (3) *arura* from ἄρουρα "land".

Over the course of its long history, Latin lexicography has produced a plethora of lexical resources, notably dictionaries, thesauri and lexica. Many are available in machine-readable form but their differing annotation schemes and formats are seldom interoperable. In an effort to offset the issue, the *LiLa: Linking Latin* project is leveraging Linked Data technology to dovetail a wide range of Latin resources into an interoperable whole, producing an ever-growing lexically-based data model capable of accommodating etymological, morphological, syntactic and semantic informa-

[1] "Captive Greece captured her savage conqueror" (our translation).

[2] In Egypt, for instance.

tion, and more besides (Passarotti et al., 2020)[3]. In LiLa, glossaries, lexica, treebanks, textual resources and tools intersect and interact through their common denominator, the lemma (itself, incidentally, a loanword from the Ancient Greek λῆμμα). Indeed, the LiLa Knowledge Base hinges on a lemma bank of approximately 130, 000 lemmas largely derived from the lexical basis of LEMLAT (Passarotti et al., 2017). As textual and lexical resources are added to the Knowledge Base, LiLa's lemma bank and coverage of the Latin lexicon grow in size.

Though chiefly targeting readily available lemmatised resources on the web, LiLa also creates linguistic resources in-house as a means of further developing its underlying data model. Examples of these are the *Index Thomisticus* Treebank (Passarotti, 2019) and *Latin VALLEX* (Passarotti et al., 2016). Here, we describe the addition of a new homegrown lexical resource, the *Index Graecorum Vocabulorum in Linguam Latinam* (Saalfeld, 1874), to the LiLa Knowledge Base of Linguistic Resources for Latin.

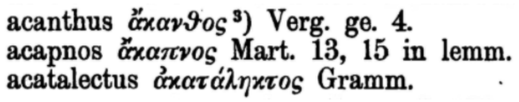
2 Data and Methodology

Etymological data is not new to LiLa. Mambrini and Passarotti (2020) describe the inclusion of 1, 391 entries from the *Etymological Dictionary of Latin and the other Italic Languages* (De Vaan, 2008) modelled against the *lemonETY* etymological extension (Khan, 2018) of the *OntoLex Lexicon Model for Ontologies (lemon)* (McCrae et al., 2017), which have provided LiLa with 1, 465 Proto-Italic and 1, 393 Proto-Indo-European reconstructed forms. Whereas those entries came to Latin via inheritance, the work described here targets (nativised) loans from Ancient Greek[4].

The *Index Graecorum vocabulorum in linguam Latinam translatorum quaestiunculis auctus* (hereafter *IGVLL*) is a list of 1, 763 Ancient Greek loanwords in the Latin language published in 1874 by classical scholar Günther Alexander E. A. Saalfeld. An extended edition of the *Index*, published in 1884 as *Tensaurus Italograecus: Ausfürliches historisch-kritisches Wörterbuch der Griechiscen Lehn- und Fremdwörter im Latenischen*, is the most comprehensive lexicographic collection of its kind, counting roughly six to eight thousand entries (Saalfeld, 1884)[5].

Of the two, the size and Optical Character Recognition (OCR) quality of the 1874 edition best suited a first development of a derivative linguistic resource, conducted as part of a Master's internship at the CIRCSE Research Centre in Milan[6].

IGVLL is structured into three columns of information: the Latin loanword (occasionally accompanied by variants), the Ancient Greek source lemma(s) (multiple lemmas include graphical, morphological and dialectal variants), and a record of attestations (see Figure 1). Explanatory notes at the bottom of the page provide additional context. In thirteen cases, question marks indicate some level of uncertainty[7], and, as is convention, asterisks are used to identify thirty-nine unattested –and thus reconstructed– Ancient Greek forms.

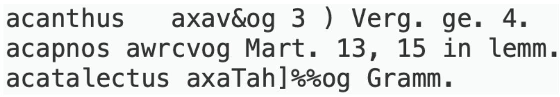


acanthus ἄκανθος ³) Verg. ge. 4.
acapnos ἄκαπνος Mart. 13, 15 in lemm.
acatalectus ἀκατάληκτος Gramm.

Figure 1: Three lexical entries in IGVLL, translating to "Bear's Foot (plant)", "without smoke", and "acatalectic (line of verse)", respectively.

2.1 Data Preparation

Judging by the illegible Greek, the engine used to produce the OCR'd text available from Internet Archive (ABBYY FineReader 8.0) was set to recognise the Latin alphabet only (see Figure 2).



```
acanthus   axav&og 3 ) Verg. ge. 4.
acapnos awrcvog Mart. 13, 15 in lemm.
acatalectus axaTah]%%og Gramm.
```

Figure 2: Latin OCR of Fig. 1, Internet Archive.

The OCR quality of the text written in the Latin alphabet, however, was sufficient to automatically isolate and tabulate the Latin lemmas, which were then manually cleaned. Next, this list was automatically mapped against the LiLa lemma bank to measure the degree of lexical overlap, which came up at 1, 488 unique matches (84.40%), 207

[3] https://lila-erc.eu/

[4] The Latin verb *græcissō* used in the title of this paper is a nativised version of the Greek γραικιζω "to imitate the Greeks; speak Greek". For a detailed overview of linguistic "nativisation", see Hock and Joseph (2009, pp. 247-57).

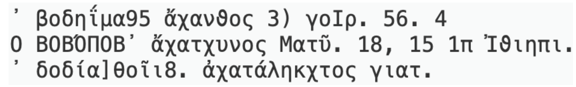
[5] Crude estimate of an average ten to fifteen entries per page, for a total 592 pages.

[6] https://centridiricerca.unicatt.it/circse_index.html

[7] The specific uncertainty remains unclear as no editorial documentation is provided.

ambiguous matches (11.74%) and 68 unmatched lemmas (3.85%). Unique matches inherited their respective LiLa identifier, ambiguous matches were manually disambiguated, and unmatched lemmas were added –once again, manually– to the LiLa lemma bank. Ambiguities were caused by homography between lemmas belonging to different categories, be those morphosyntactic (the lemma *philosophus*, for instance, matched against LiLa's adjective *philosophus* "philosophical" and common noun *philosophus* "philosopher") or inflectional (the common noun *er* might refer to the masculine *er, eris* "hedgehog" or the invariable *er* (graphical variant of *R*) "seventeenth letter of the Latin alphabet"). Of the 68 unmatched lemmas, 33 were graphical variants of lemmas already present in LiLa and 35 were new additions.

Next, we OCR'd IGVLL with Tesseract v. 4.1.1 set to Ancient Greek recognition (Smith, 2007)[8]. As Figure 3 shows, contrary to the Latin OCR the noise affecting Greek lemmas required heavy manual intervention for clean tabulation, e.g. the rectification of instances of ϰ (cappa) misread as χ (chi) or of π (pi) misread as ττ (double tau) and viceversa, missing breathings and incorrect accents, to mention but a few.



```
’ βοδηΐ́μα95 ἄχανθος 3) γοΙρ. 56. 4
O BOBΌΠOB᾿ ἄχατχυνος Ματῦ. 18, 15 1π Ἰθιηπι.
’ δοδία]θοΐι8. ἀχατάληκχτος γιατ.
```

Figure 3: Ancient Greek OCR of Fig. 1, Tesseract.

In LiLa, a lemma can have one or more graphical variants, known as "written representations" (e.g. the verb *sacrifico* "to sacrifice" is also attested as *sacrufico*), as well as inflectional variants, with which it holds a symmetric "lemma variant" property or relation in the Knowledge Base (the active *sacrifico, sacrufico* vs. the deponent *sacrificor, sacruficor*).

Therefore, for the purposes of LiLa, where the editor provides multiple Ancient Greek lemmas for a single Latin loanword, e.g. *burrus* "red" πυρρός (πυρσός); *cyperum* "rush (botany)" κύπειρον (κύπειρος), these were distinguished into written representations of the same lemma (i.e. πυρρός vs. πυρσός) and lemma variants (i.e. the neuter κύπειρον vs. the masculine κύπειρος).

Compounds such as *authepsa* "an urn, boiler", (αὐτός & ἕψω) were tabulated as two separate words, and entries followed by a question mark (13 in total) were marked as "uncertain".

2.2 Data Model

The transformation of IGVLL into an RDF lexicographic resource bound for LiLa relied on a combination of vocabularies. In line with previous etymological work, we integrated the aforementioned lemon and lemonETY modules of OntoLex to represent lexical entries in IGVLL. The example lemma *abacus* "sideboard" shown in Listing 1 is treated as an `ontolex:LexicalEntry` linked to LiLa's own *abacus* (lemma ID `86829`) through the property `ontolex:canonicalForm`.

```
a                            ontolex:LexicalEntry;
rdfs:label                               "abacus";
ontolex:canonicalForm <..lemma/86829>;
```

Listing 1: Latin

We employed the *Simple Knowledge Organization System* (*SKOS*) (Miles and Bechhofer, 2009) to point Ancient Greek lemmas to their corresponding canonical forms in a machine-readable version of the Greek-English *Liddell-Scott Jones (LSJ)* lexicon (Blackwell, 2018). As Listing 2 shows, we modelled the Ancient Greek source lemma of *abacus*, ἄβαξ, as an `etymon`, which, in the absence of a Linked Data Knowledge Base for Ancient Greek, currently points to a blank node.

```
a                                 lemonEty:etymon;
rdfs:label                                "᾿άβαξ";
lime:language                              "grc";
ontolex:canonicalForm
           [ontolex:writtenRep "᾿άβαξ"];
skos:exactMatch    <urn:cite2...:n51>.
```

Listing 2: Ancient Greek

The `skos` property stores the LSJ identifier of ἄβαξ as an `exactMatch` to denote an exact correspondence between the Ancient Greek lemma of IGVLL and that of LSJ. Failing an exact match, the property `skos:broadMatch` is used to indicate that the IGVLL lemma is incorporated in a different entry of LSJ (e.g. the IGVLL noun φυσική "science of nature, physics" does not have its own entry in LSJ but is listed as a nominalised adjective under the adjectival entry φυσικός "natural"); further, failing both exact and broad matches, the property `skos:relatedMatch` is used to indicate a loose relation between IGVLL and LSJ (e.g. IGVLL's πορφυρίζον "purple

[8] For the most recent overview of Ancient Greek optical character recognition, see Robertson and Boschetti (2017).

dye pigment", neuter present participle of πορφυρίζειν, and LSJ's verb πορφυρίζω "to be purplish"). As LSJ is not currently equipped with a URN resolver, no actionable link can be made between LiLa and LSJ.

If multiple written representations of a Greek word are listed in the IGVLL, those are all assigned to the canonical form of the related `etymon`, for instance `ontolex:canonicalForm [ ontolex:writtenRep "πυρρός", "πυρσός" ].`

In the case of multiple Ancient Greek variant lemmas, these are all treated as individual etyma, with the difference that the primary `etymon` points to the URI(s) of the other etyma –classed as both `lemonEty:etymon` and `lemonEty:cognate`– via the additional property `lemonEty:cognate` (Listing 3).

```
a                      lemonEty:etymon;
rdfs:label                  "κύπειρον";
lime:language                    "grc";
ontolex:canonicalForm
     [ ontolex:writtenRep "κύπειρον" ];
skos:exactMatch <urn:cite2...n60988>;
lemonEty:cognate
 <http://lila.../IGVLL/id/etymon/499>.
```

Listing 3: Lemma variants: κύπειρον/ος

Latin composite words in IGVLL never point to an Ancient Greek compound but to the two constituent lemmas. In contrast, in the LSJ lexicon seven of the total thirteen multi-word lexical entries in IGVLL are traced back to a Greek compound lemma, e.g. *authepsa* (IGVLL: αὐτός & ἕψω; LSJ: αὐθέψης[9]). In keeping with the IGVLL, we employed the `decomp:subterm` property of *lemon*[10] to point the Latin lexical entry to its two constituent Ancient Greek etyma and reconciled these with LSJ using the `skos:relatedMatch` property.

The etymology of *abacus* is expressed with the *CIDOC Conceptual Reference Model (CRM)* class `E89 Propositional Object`[11] as a borrowing by way of the `lemonEty:etyLinkType` property. This set-up is also valid for calques, should these become available in future.

The *CRM$_{inf}$* extension of CRM and the *Open Vocabulary* (Davis, 2004) were used to represent uncertainty as a "belief" or confidence value (Stead et al., 2019; Doerr, 2003; Mambrini and Passarotti, 2020). Specifically, we coded uncertainty as a CRM$_{\text{inf}}$ `Belief` class (`crminf:I2`) carrying an arbitrary `Belief Value` (`crminf:I6`) of `0.5` (Listing 4).

```
a                          crminf:I2;
crminf:J5              [a crminf:I6;
                 ov:confidence 0.5].
```

Listing 4: Uncertainty

Additionally, we employed the *Dublin Core*^{TM} *Metadata Terms* vocabulary to supply the resource with descriptive metadata, such as publisher and licence (DCMI, 2020).

All editorial notes in IGVLL were excluded from the data model.

As previously mentioned, with this development LiLa's etymological purview now covers both direct inheritance *and* borrowing. Figure 4, for example, shows all etymological information in the Knowledge Base associated with LiLa's common noun *muscus* "moss, musk" (top row, centre node). LiLa's "muscus" is connected to the "muscus" lexical entries of both IGVLL and the Brill Etymological Dictionary via the bidirectional OntoLex property `canonicalForm`. These lexical entries point to their respective etyma via the directed lemonETY `etymology` and `etymon` properties.

3 Conclusion

This paper describes the preparation and integration of Saalfeld's *Index Graecorum Vocabulorum in Linguam Latinam* (1874) in the LiLa Knowledge Base of Linguistic Resources for Latin. This first list of 1,763 Latin loans from Ancient Greek adds 68 new Latin lemmas to LiLa, stretches its data model to include borrowing and has been mapped to the digitised Greek-English Liddell-Scott-Jones lexicon. Beyond LiLa, this linguistic resource might be integrated in other resources, such as dictionaries (Bowers and Romary, 2016) or digital scholarly editions. Future improvements might acquire a list of calques (Detreville, 2015; Fruyt, 2011) and the extended edition of Saalfeld's *Index* (1884).

The data and code for the project are available at: `https://github.com/CIRCSE/index-graecorum-vocabulorum`.

[9] `http://folio2.furman.edu/lsj/?urn=urn:cite2:hmt:lsj.chicago_md:n17373`

[10] `https://lemon-model.net/lemon-cookbook/node21.html`

[11] `http://www.cidoc-crm.org/Entity/e89-propositional-object/version-6.0`

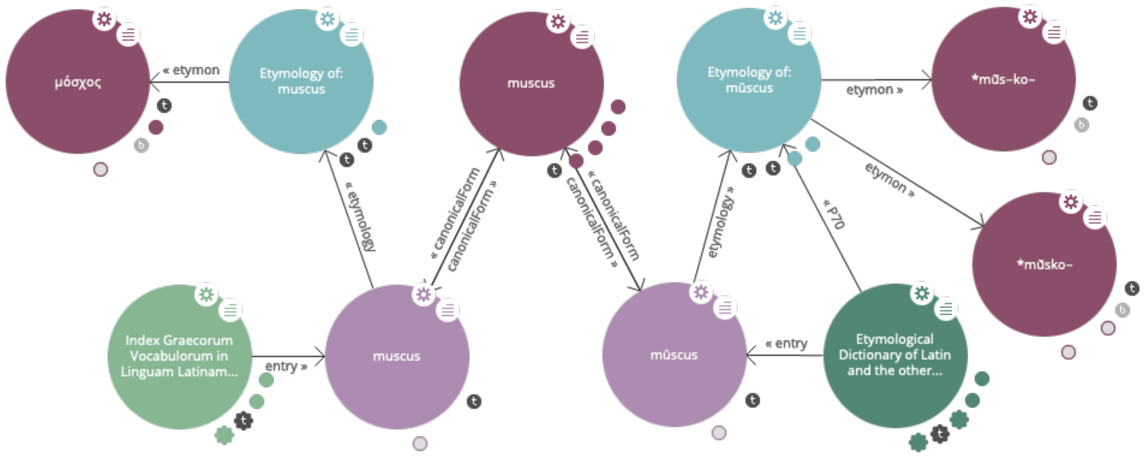


Figure 4: Etymology of *muscus* “moss, musk” in LiLa.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme - Grant Agreement No 769994.

References

James Noel Adams. 2003. *Bilingualism and the Latin Language*. Cambridge University Press.

Christopher W. Blackwell. 2018. CITE LSJ Browser, v. 1.4.0. `http://folio2.furman.edu/lsj/index.html`.

Jack Bowers and Laurent Romary. 2016. Deep Encoding of Etymological Information in TEI. *Journal of the Text Encoding Initiative*. `https://doi.org/10.4000/jtei.1643`.

Ian Davis. 2004. vocab.org, v. 20040205. `https://vocab.org/`.

Usage Board DCMI. 2020. DCMI Metadata Terms. `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`.

Michiel De Vaan. 2008. *Etymological Dictionary of Latin: And the other Italic Languages*, volume 7 of *Leiden Indo-European Etymological Dictionary Series*. Brill, Amsterdam. `https://brill.com/view/title/12612?language=en`.

Eleanor Detreville. 2015. An Overview of Latin Morphological Calques on Greek Technical Terms: Formation and Success. Master’s thesis, University of North Carolina, Asheville, NC.

Martin Doerr. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75. `https://dl.acm.org/doi/10.5555/958671.958678`.

Michèle Fruyt. 2011. Latin Vocabulary. In James Clackson, editor, *A Companion to the Latin Language*, pages 144–156. Wiley-Blackwell.

Hans Henrich Hock and Brian D. Joseph. 2009. *Language history, language change, and language relationship: an introduction to historical and comparative linguistics*. Trends in Linguistics. Studies and monographs, 218. Mouton de Gruyter, 2nd revised edition. `http://gen.lib.rus.ec/book/index.php?md5=819682013cda444e5dad7bf866a45d64`.

Anas Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9:304–320. `https://doi.org/10.3390/info9120304`.

Adam Ledgeway. 2012. *From Latin to Romance. Morphosyntactic Typology and Change*. Oxford University Press, Oxford.

Francesco Mambrini and Marco Passarotti. 2020. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi, and Besim Kabashi, editors, *Proceedings of the Globalex Workshop on Linked Lexicography (LREC 2020)*, pages 20–28, Paris, France. European Language Resources Association (ELRA). `https://doi.org/10.5281/zenodo.3862156`.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of the Electronic lexicography in the 21st century conference (eLex 2017)*, pages 237–251. `https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf`.

Alistair Miles and Sean Bechhofer. 2009. SKOS Simple Knowledge Organization System Reference. *W3C recommendation*, 18. `https://www.w3.org/TR/skos-reference/#L4858`.

Marco Passarotti, Berta González Saavedra, and Christophe Onambele. 2016. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2599–2606. https://www.aclweb.org/anthology/L16-1414/.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31. http://www.ep.liu.se/ecp/article.asp?issue=133&article=006&volume=.

Marco Passarotti, Francesco Mambrini, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas: The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58:177–212. https://doi.org/10.4454/ssl.v58i1.277.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, volume 10 of *Age of Access? Grundfragen der Informationsgesellschaft*, pages 299–319, Berlin-Boston. De Gruyter GmbH. https://doi.org/10.1515/9783110599572.

Bruce Roberston and Federico Boschetti. 2017. Large-Scale Optical Character Recognition of Ancient Greek. *Mouseion*, 14:341–359. https://doi.org/10.3138/mous.14.3-3.

Günther Alexander E. A. Saalfeld. 1874. *Index graecorvm vocabvlorvm in lingvam latinam translatorum qvaestivncvlis avctvs*. apvd F. Berggold, Berlin. https://archive.org/details/indexgraecorvmvo00saal/.

Günther Alexander E. A. Saalfeld. 1884. *Tensaurus Italograecus: Ausführliches historisch-kritisches Wörterbuch der griechischen Lehn- und Fremdwörter im Lateinischen*. Carl Gerold's Sohn, Wien. https://archive.org/details/tensaurusitalogr00saal.

Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. https://doi.org/10.1109/ICDAR.2007.4376991.

Stephen Stead, Martin Doerr, Christian-Emil Ore, and Athina et al. Kritsotaki. 2019. CRMinf: the Argumentation Model, Version 0.10.1 (draft). http://new.cidoc-crm.org/crminf/sites/default/files/CRMinf\%20ver\%2010.1.pdf.