Linking the Computational Historical Semantics corpus to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Giulia Pedonese and **Flavio Massimiliano Cecchini** and **Marco Passarotti**
Università Cattolica del Sacro Cuore, Italy
giulia.pedonese@unicatt.it
flavio.cecchini@unicatt.it
marco.passarotti@unicatt.it

Abstract

This paper describes the linking of a subset of five texts from the Latin Text Archive corpus of the Computational Historical Semantics project to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin for a total of about one million tokens, adding approximately 13 million and 750 thousand new triples to the Knowledge Base. To show the potentialities of linking those texts to other resources for Latin, the paper describes the results of a sample query conducted on the texts linked to the Knowledge Base.

1 Introduction and related work

Thanks to its key role in accessing the European cultural heritage, Latin was one of the first languages to be automatically processed. Since the pioneering work of the late Fr. Roberto Busa SJ on Thomas Aquinas' texts in 1949 (Nyhan and Passarotti, 2019), an abundance of linguistic resources has been made available for Latin as a result of a long tradition of studies in the area of Computational Linguistics, Literary Computing and Digital Humanities. These include textual resources such as corpora featuring texts of various typologies, as well as lexical resources such as lexica, dictionaries and thesauri. Besides larger (meta)collections of texts such as the *Corpus Corporum*,[1] which contains more than 150 million words provided by more than twenty different collections, among the corpora providing more specific data there are, for example, the *Patrologia Latina* data base,[2] featuring the writings of the Church Fathers, and the *Musisque Deoque* digital archive, which contains poetic works from Classical to Late Latin.[3] Lexical resources include the *Thesaurus Linguae Latinae* at the Bayerische Akademie der Wissenschaften in Munich,[4] Johann Ramminger's *Neulateinische Wortliste*,[5] and Lewis and Short's dictionary (Lewis and Short, 1879), accessible among others through the Perseus Digital Library and now linked to the LiLa Knowledge Base (Mambrini et al., 2021).

Unfortunately, while there is a large number of linguistic resources for Latin currently available in digital format, these often lie scattered in isolated "data silos", a fact which prevents users from exploiting their full potential in interoperable ways: linguistic data and metadata for Latin are distributed in separate collections which often use different data formats, query languages, annotation criteria and tagsets, thus making the resources incompatible with each other. In the last decade, multiple efforts have been made to provide a solution to the problem of dispersion of (meta)data and resource isolation. Today, many initiatives offer a single access point to resources collected in single repositories, such as the European infrastructure CLARIN,[6] the metadictionary *Logeion*,[7] and the already mentioned metacollection *Corpus Corporum*. However, such initiatives still fail to provide real interoperability between distributed linguistic resources, which would require "that all types of annotation applied to a particular word/text be integrated into a common representation for indiscriminate access to any linguistic information provided by a resource or tool" (Chiarcos, 2012a, p. 162). A current approach to interlinking linguistic resources is that of the Linguistic Linked Open Data cloud, a collaborative effort pursued by several members of the Open Linguistics Working Group[8] with the goal of applying the Linked Data principles to linguistic data.[9]

[1] https://www.mlat.uzh.ch/
[2] https://www.lib.uchicago.edu/efts/PLD/
[3] https://mizar.unive.it/mqdq/public/
[4] https://tll.degruyter.com/
[5] http://nlw.renaessancestudier.org/
[6] https://www.clarin.eu/
[7] https://logeion.uchicago.edu/
[8] http://linguistic-lod.org/llod-cloud
[9] Among the initiatives combining the Linked Data technologies and language resources is the COST action *Nexus Lin-*

The Linked Data paradigm consists of a series of best practices and principles for exposing, sharing and connecting data on the web, which are incarnated by the following rules:[10]

- data and metadata should be unequivocally named by URIs (Uniform Resource Identifiers), allowing users to find them;
- HTTP URIs should be used in order for data to be accessible by both humans and machines;
- provide useful information through Web standards such as the RDF data model (i. e. Resource Description Framework), which represents data in the form of triples: a predicate property (1) connecting a resource called subject (2) to another resource, called object (3). In this way, data are represented through directed, labelled graphs and are searchable via another Web standard like the SPARQL query language (the language used to query data in RDF format);
- include links to other URIs in order to allow for further research.

Applying the Linked Data paradigm is a way to share data according to the FAIR principles, which state that data must be Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). The LiLa Knowledge Base of linguistic resources for Latin aims to make textual and lexical resources interoperable trough the application of the Linked Data principles (see Section 2).

After introducing the architecture of the LiLa Knowledge Base (Section 2) and the Computational Historical Semantics project (Section 3), this paper describes the linking to LiLa of a textual resource consisting of Medieval documentary Latin texts taken from the Latin Text Archive of the Computational Historical Semantics project (Section 4). Finally, the paper provides an example of query to show the potentialities of interlinking those texts to other resources for Latin (Section 5) and gives insights into the future developments of LiLa (Section 6).

guarum, whose aim "is to promote synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science" (at https://nexuslinguarum.eu/the-action/, *What the Action does*).

[10] https://www.w3.org/DesignIssues/LinkedData

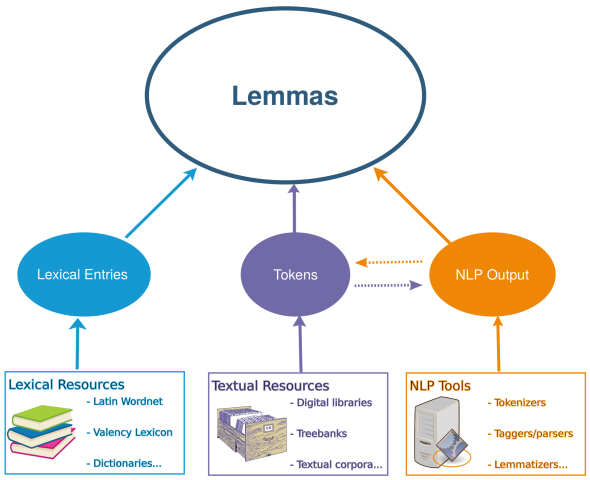

Figure 1: The architecture of the LiLa Knowledge Base.

2 The LiLa Knowledge Base

The *LiLa - Linking Latin* project[11] aims to connect the existing linguistic resources for Latin in order to make them interoperable (Passarotti et al., 2020). The LiLa team is building an open-ended Knowledge Base following a set of standards for the Semantic Web and Linked Data. To this end, all content involved or referenced in the linguistic resources connected in LiLa is made unambiguously findable and accessible by assigning each data point an HTTP URI. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as the RDF data model (see Section 1) and the SPARQL query language.[12] Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies in order to model relevant information. The main ones are: POWLA for corpus data (Chiarcos, 2012b), OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015), and Ontolex-Lemon for lexical data (Buitelaar et al., 2011; McCrae et al., 2017).

Within this framework, LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and Natural Language Processing (NLP) tools. Consequently, the architecture of the LiLa Knowledge Base is highly lexically-based (cf. Figure 1), being grounded on a simple but effective assumption that strikes a good balance between feasibility and granularity: Tex-

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

[12] LiLa's SPARQL endpoint can be accessed at: https://lila-erc.eu/sparql/

tual resources are made of (occurrences of) words (more precisely, *tokens*), lexical resources describe properties of words (in *lexical entries*), and NLP tools process words (producing NLP *outputs*).[13]

Considering the central role played by lemmas in LiLa, the core of the knowledge base is the so-called *Lemma Bank*,[14] a collection of about 200 000 Latin lemmas (defined as the canonical forms of lexical items, i. e. their citation forms) originally taken from the data base of the morphological analyzer LEMLAT (Passarotti et al., 2017). Interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. The resources currently linked to the knowledge base are as follows:

- **Textual resources**
  - Computational Historical Semantics: 1 058 084 tokens
  - *Confessiones*: 92 351 tokens
  - Corpus for Latin Sociolinguistic Studies on Epigraphic texts: 32 473 tokens
  - Index Thomisticus Treebank: 450 515 tokens
  - LASLA corpus: 1 839 373 tokens
  - *Liber Abbaci* (ch. VIII): 29 858 tokens
  - *Querolus sive Aulularia*: 13 232 tokens
  - UDante Treebank: 55 287 tokens
- **Lexical resources**
  - Lemma Bank: 153 965 entries
  - Etymological Dictionary of Latin and the other Italic Languages: 1 452 entries
  - Glossary of Latin loanwords from the Italian works of Dante Alighieri: 765 entries
  - Index Graecorum Vocabulorum in Linguam Latinam Translatorum: 1 759 entries
  - LatinAffectus: 3 295 entries
  - Latin Vallex 2.0: 3 561 entries
  - Latin WordNet: 6 269 entries
  - Lewis & Short's dictionary: 53 437 entries
  - Word Formation Latin: 41 791 entries

As shown in Section 3, the subset of the Computational Historical Semantic corpus adds a significant amount of Late and Medieval Latin texts, expanding the possibilities of integrated research with other (Medieval) Latin corpora such as the Index Thomisticus Treebank and UDante.

3 Computational Historical Semantics

Computational Historical Semantics (from now on CompHistSem) is a co-operative project involving the German universities of Bielefeld, Frankfurt am Main, Regensburg and Tübingen, originally developed by an interdisciplinary team led by Bernhard Jussen and Alexander Mehler at the Goethe University in Frankfurt am Main, and funded by the German Federal Ministry for Education and Research.[15] The project aims to define new methods and tools for historical-semantic analysis "by conducting computer-based research on processes of linguistic change" (Cimino et al., 2015).

The associated website[16] of the *Latin Text Archive* (LTA), hosted by the Berlin-Brandenburg Academy of Sciences and Humanities, allows users to simplify their search for semantic and linguistic changes by quickly comparing a large number of texts gathered from various sources: more than 4 000 texts spanning from the 2nd to the 15th Century AD, put together thanks to the support of digitalised collections such as the *Patrologia Latina* data base, the *Monumenta Germaniae Historica* (MGH),[17] the *Corpus Corporum* (University of Zürich) and the *Bibliotheca Augustana*.[18] These texts are lemmatised by means of the Frankfurt Latin Lexicon (FLL), a morphological lexicon of Medieval Latin organised around three "lexical resolutions" of lexical units (Mehler et al., 2020) which enable a multilayered search:

1. the *superlemma*, providing a unified representation for different variants of a "word" (i. e. a lexeme), e. g. *caelum* 'sky', as opposed to

2. *lemmas*, which are tied to specific variants of a word, e. g. *cael*, *caelum*, *cælum*, *caelus*, *celum*, *cęlum*, *celus*, *coelum*, *cœlum*, *coelus*, each with its own spelling and possibly inflected according to different paradigms, which consist of

[13] In Figure 1, the arrows going from and to the node for NLP *Output* represent the fact that tokens that are the outputs of a specific NLP tool (a tokeniser) can become the inputs of further tools (like, for instance, a syntactic parser).

[14] http://lila-erc.eu/lodview/data/id/lemma/LemmaBank

[15] https://comphistsem.org/home.html. NB: this site is no longer maintained.

[16] https://lta.bbaw.de/

[17] https://www.mgh.de/

[18] http://www.hs-augsburg.de/~harsch/augustana.html

3. *word forms*, such as *cęlorvm* (lemma *cęlum*) or *coelos* (lemma *coelus*), possibly tagged for morphological features such as *casus* (case) or *numerus* (number).

While the FLL allows a user to search for a specific word or word form and obtain quantitative data with respect to its occurrences as well as grammatical, linguistic and lexical information about its use, the textual data base LTA makes it possible to perform a text-based search of the whole corpus, and is useful to carry out more complex searches for word co-occurrences (Cimino et al., 2015). Since CompHistSem is an ongoing project, it is constantly expanding as more texts, words and word forms are added to its data bases (Mehler et al., 2020).

4 Linking *CompHistSem* to LiLa

In this section, the process adopted so as to link texts from the CompHistSem project to the LiLa Knowledge Base is detailed: first in general, and then by giving a more in-depth discussion of problematic cases.

4.1 Texts, annotation and format conversion

The linking procedure is implemented on a subset of the LTA corpus of CompHistSem consisting of seven texts or text collections. These are the texts that have been selected by the CompHistSem team after having been requested for data from their corpus to include into LiLa, and that have been deemed of sufficient size for this goal. The specific documents are:

- *Capitularia Regum Francorum*, 6th–9th c. AD, various authors, from MGH Capitularia 1 & 2
  - 10 820 sentences,[19] 343 030 tokens (including 53 161 punctuation marks)
- *De ecclesiasticis officiis*, 9h c. AD, by Amalarius of Metz, from Patrologia Latina vol. 105
  - 4 279 sentences, 125 475 tokens (including 20 845 punctuation marks)
- *Vita Karoli Imperatoris*, 9th c. AD, by Einhard, from MGH Scriptores rerum Germanicarum 25
  - 247 sentences, 8 393 tokens (including 1 224 punctuation marks)
- *Gesta Hludowici imperatoris*, 9th c. AD, by Thegan of Trier, from MGH Scriptores rerum Germanicarum 64
  - 451 sentences, 8 355 tokens (including 1 403 punctuation marks)
- *Decretum Gratiani* I to III (treated as distinct documents), also known as *Concordia discordantium canonum*, 12th c. AD, by Gratian, from Corpus Corporum through Patrologia Latina vol. 187
  - 31 803 sentences, 572 831 tokens (including 124 656 punctuation marks)

In total, there are 47 600 sentences for 1 058 084 tokens (including 201 289 punctuation marks), the vast majority of which (see Section 4.2) lemmatised and tagged for parts of speech and morphological features by means of the Frankfurt Latin Lexicon (see Section 3), which uses its own tagset, in line with the grammatical categories traditionally recognised for Latin.[20] All texts but the *Decretum Gratiani* (Corpus Corporum, transcription under Creative Commons Share-Alike license[21]) are retrievable from the LTA (see Section 3) and are under the Creative Commons license.[22] The texts are encoded in the TEI-P5 format, i. e. as XMLs.[23]

The preliminary step before linkage is the conversion of the XMLs to the CoNLL-U format,[24] as used in the Universal Dependencies (UD) project (de Marneffe et al., 2021), by means of a Python[25] script developed as part of the LiLa project's endeavour.[26] The motivation for this move is twofold: first, the CoNLL-U format is more easily human-readable, with no loss of information nor of machine-readability with respect to the original XML; second, the conversion of format also entails a conversion of part-of-speech and morphological tags, similarly to what has already been achieved for other data sets, such as the Index Thomisticus Treebank (Cecchini et al., 2018) or the Late Latin

[19] "Sentence" in this context refers to the textual segmentation inherited from CompHistSem, and does not necessarily coincide with a syntactically-driven interpretation thereof; this however is irrelevant here, as only single tokens are considered.

[20] A classic and accessible reference for Latin is (Greenough et al., 2014).

[21] https://creativecommons.org/licenses/by-sa/4.0/

[22] https://creativecommons.org/licenses/by/4.0/

[23] https://tei-c.org/

[24] https://universaldependencies.org/format.html

[25] www.python.org

[26] The script has not yet been made public.

Charter Treebank (Cecchini et al., 2020a). The latter point is relevant, since also LiLa makes use of UD's part-of-speech tagset internally, and so the conversion to the CoNLL-U format has the ultimate effect of better integrating CompHistSem texts into the knowledge base and of laying the ground for its linking, at the same time acting as a stepping stone towards a possible future annotation according to UD guidelines.

The mapping between the two tagsets is rather straightforward, especially with regard to morphological tags, whose distribution already broadly corresponds to that found in the UD formalism applied to Latin, or can be implemented on a lexical basis. Parts of speech also overlap or are retraceable to more general classes (e. g. CompHistSem's distributives DIST and ordinals ORD merge into UD's adjectives ADJ with a corresponding value of the NumType feature[27]) to a great degree, since they have common roots in traditional grammars, but need some further reworking: in particular, the class of determiners (in UD labeled as DET) has to be carved out from CompHistSem's adjectives (ADJ) and pronouns (PRO); a difference has to be drawn, on a lexical basis, between co-ordinating (CCONJ in UD) and subordinating (SCONJ) conjunctions; some readjustments between indeclinable classes (especially adverbs, ADV in UD; conjunctions, CCONJ/SCONJ; particles, PART) are necessary; and tokens with atypical lemmas such as *biblical books* and/or belonging to mixed nominal or residual classes (Noun, NE, NP, PTC, XY, FM in CompHistSem) require some case-by-case treatment.

4.2 Lemmatisation

Since LiLa is structured around the notion of lemma (see Section 2), which is the key element through which lexical and textual resources are connected to the knowledge base, lemmatisation of a document is a necessary step in order to proceed with the linking process. As mentioned in Section 4.1, this is already the case for texts found in the LTA: the LEMMA field in the CoNLL-U conversion (see Section 4.1) directly stores the *superlemma* relative to the word form, as determined per the Frankfurt Latin Lexicon (see Section 3).

Only a negligible 2 697 tokens lacking a lemma are detected, i. e. the 0,25% of the total, for which the Frankfurt Latin Lexicon fails to produce one. They represent 1 775 (case-sensitive) form types, and mostly consist of proper nouns, or terms derived from proper nouns (hence conventionally capitalised), such as *Magonciam* 'Mainz (city in Germany)', variant of a more Classical *Mogontiacum*, or *Tolletano* 'from Toledo (city in Spain, *Toletum* in Latin)', but also forms such as *ſ* or *ſſ*. Given the peculiar, onomatological nature and marginality of such forms, and the fact that in this phase the focus is on linking and not on expanding LiLa's lexical data base, these tokens are not considered further and left out from lemmatisation (and thus linking).

More in general, it has to be noticed that the data from CompHistSem, as that of any other external resource, is taken 'as is': it is not the goal nor the scope of this work to assess the "correctness" of any level of its annotation (tokenisation, lemmatisation, part-of-speech-tagging, morphological features). The aim here is only to link different resources to the LiLa Knowledge Base, without intervening in their annotation standards: this means that no evaluation is performed, nor can be, as LiLa itself avoids establishing a standard. However, the interoperability of many different resources can surely help achieve an overview of the variations between annotation formalisms, in view of a possible harmonisation of their criteria, e. g. in a typological framework (cf. Gamba and Zeman 2023).

4.3 Matching and non-matching tokens

Even if no evaluation in a true sense can be performed, the complexity of the linking task can be gauged by looking at the different cases that present themselves and at the strategies that are necessary to deal with them, and how they are distributed among the tokens. First and foremost, the trivial case of punctuation marks is ignored: besides being invariably assigned a lemma identical to their form and part of speech PUNCT, and thus not presenting any ambiguity, punctuation marks are not lexical units, and as such do not even appear in the LiLa lemma bank. This brings it down to 856 795 "lexical tokens"[28] that can be contemplated for linking from the original total of 1 058 084. In the following, a breakdown of the outcomes of the linking

[27] We point to UD guidelines, which can be browsed at https://universaldependencies.org/guidelines.html, for details about the meaning of labels in the UD framework.

[28] "Lexical" in the sense of corresponding to what is usually considered to be a word (with all its indefiniteness, cf. Haspelmath 2017), not necessarily as in the lexical/functional dichotomy of UD (see de Marneffe et al., 2021, §2.1.1).

process is given, at the end of which approximately 13 million 750 thousand new triples are added to the LiLa Knowledge Base.

4.3.1 Unambiguous matches

As many as 720 860 of these lexical tokens can be directly linked to the LiLa knowledge base through an unambiguous match in the LiLa lemma bank with their respective combinations of lemma and part of speech (after conversion, see Section 4.1): an example is the lemma *itinerarium* 'itinerary' coupled with the part of speech NOUN, a combination which exists and is unique in LiLa.[29] It has to be remarked that such a match is independent from the specific word form: this is the advantage of pivoting on the (super)lemma, as it abstracts from not always predictable spelling and inflection variants. The total coverage of direct linking is thus the 84,14% of all tokens; if only the number, 18 262, of unique combinations of lemma and part of speech among lexical tokens in our subcorpus is taken into account, the coverage is instead 68,50% (12 509 combinations). This difference arises from the fact that many unambiguously linked tokens represent very frequent functional words such as the co-ordinating conjunction (CCONJ) *et* 'and' (33 250 occurrences) or the pronoun (PRON) *qui* 'who, which, that' (17 434 occurrences), while the vocabulary of the chosen texts indeed sensibly departs from the original lexical pool of the LiLa lemma bank (cf. Section 5).

Again, it has to be noticed that no upstream control is performed on the criteria or correctness of the lemmatisation in CompHistSem: all the just described unambiguous matches are inserted as they are, meaning that, in a sense, LiLa accepts the risk of picking up spurious forms.

4.3.2 Ambiguous matches

There are cases in which a token's combination of lemma and part of speech can be matched to more than one entry in the LiLa lemma bank: in particular, this happens for 54 903 lexical tokens (corresponding to 777 lemma/part-of-speech types), e. g. for the lemma *contingo* 'to touch' or 'to wet' coupled with the part of speech VERB, for which we have three candidates.[30] In all these cases, each token proceeds to be linked to all its suitable candidates, leaving the linking ambiguous. This is an acceptable compromise in the face of the relatively low incidence of such ambiguities, and of the fact that some tokens would still not be distinguishable even when taking into account all other morphological factors: e. g. for *contingo* VERB, knowing that its word form is *contingat* and that its mood is subjunctive, still one could not choose between entry 93415 or 96293 in the LiLa lemma bank. A contextual and/or semantic disambiguation would take an unnecessary effort and is outside the scope of the linking task presented here.

4.3.3 No matches

There are 81 032 lexical tokens left that cannot be retraced to any entry in the LiLa lemma bank. This can have three reasons:

1. either the token does not possess a lemma, or
2. it has a lemma unknown to LiLa, or finally
3. there is a mismatch between lemma and part of speech from the point of view of the LiLa lemma bank.

1. As discussed in Section 4.2, the first case is marginal, and those tokens are ignored.

2. The second case is exemplified by the lemma *subplantatio* (with part of speech NOUN): it is a regularly formed, if novel, Latin word for which it is possible to extract all necessary values to insert it in LiLa's lemma bank from CompHistSem's annotation. However, since it is not already in the lemma bank, it cannot yet be linked at this stage. The number of different types (with respect to lemma, part of speech and morphological features) of new words ready for insertion is 2 448, but if 257 with residual part of speech X (meaning they do not have a meaningful analysis from the point of view of Latin, being mostly foreign words) are discarded, together with 693 numerals expressed as digits or Roman numerals, the remaining lexical items not unexpectedly show a preponderance of 699 proper nouns (PROPN), e. g. *Teudericus*, followed by 378 adjectives (ADJ), e. g. *adrianopolitis* 'from the city of Adrianopolis (modern-day Edirne, in Turkey)', 257 common nouns (NOUN), e. g. *pyromantica* 'divination by fire' (related to the already known *pyromantia*), 45 verbs (VERB), e. g. *exonio* 'to excuse',[31] 30

[29] https://lila-erc.eu/data/id/lemma/109142

[30] http://lila-erc.eu/data/id/lemma/43870, http://lila-erc.eu/data/id/lemma/93415 and http://lila-erc.eu/data/id/lemma/96293.

[31] Cf. http://ducange.enc.sorbonne.fr/exonia.

adverbs (ADV), e. g. *nudiustertius* 'now three days ago', 6 literal numerals (NUM), e. g. *uigintiquinque* 'twenty-five', 3 pronouns (PRON), e. g. *nosipsi* 'we ourselves', 3 interjections (INTJ), e. g. *hosanna* 'hosanna, praise', and 2 subordinating conjunctions (SCONJ), e. g. *quamobrem* 'for what reason'.[32] A further 429 lemmas with a part of speech can be identified, e. g. the PROPN *Ebbo*, for which however morphological features are lacking, and for which therefore some research is needed before insertion/linkage. The distribution of all these missing lemmas, skewed towards names of persons and places, already gives an interesting picture of the character and provenance of the documents at hand, which is further explored at the phrase level in Section 5.

3. The third case is again split between those tokens having a unique possible match (with respect to their lemmas) with an entry in the LiLa lemma bank, and those having multiple possible matches. In both events, the misalignment with the corresponding parts of speech found in the LiLa lemma bank means that all these 2 426 lemma/part-of-speech types have to be manually checked to understand if there is a presence of false matches (which could eventually lead to new insertions in LiLa's lemma bank), or deviating standards of annotation. The latter case is illustrated by the rather frequent (1 606 occurrences) lemma *ita* 'thus, so' misleadingly labelled as a conjunction in CompHistSem, while it appears as an adverb (ADV) in the LiLa lemma bank. There are some "internal" misalignments, too: the negation *non* 'not' (taking up alone 16,71% of all missing matches, with 13 538 occurrences) is tagged as a particle (PART) in the CoNLL-U conversion according to UD standards,[33] but is registered as an adverb (ADV) in LiLa.

Also, the morphological analyser LEMLAT[34] (Passarotti et al., 2017) is deployed directly on word forms to check if some annotation choices in CompHistSem, unrecognised by LiLa, do fall into the category of *hypolemmas*, i. e. a standard word form that represents a well-defined subset of the inflectional paradigm of a lemma, which under some criteria might be considered to be a lemma itself: among the most common examples are participles (see below) (Passarotti et al., 2020).[35] So, for example, this strategy leads to envisage LiLa's entry of the adjective (ADJ) *caelestis*[36] 'heavenly' for what in the CompHistSem's texts is labelled as the common noun (NOUN) with lemma *caeleste*, i. e. the substantivised neutral singular form of the adjective, which would have been otherwise undetectable, as *caeleste* does not appear as an individual entry in LiLa's lemma bank. Under this light, an example of a false match that needs to be rejected is the entry NOUN *paterium*[37] 'a kind of Evangeliary'[38] for a possible proper noun *Paterius*: in fact, Paterius was the name of a bishop of Brescia in the 6th Century AD. Among misalignments, there are some recurring cases that can be treated systematically:

- misalignments between NOUNs and ADJs and vice versa, which mostly happen when a substantivised adjective is considered an independent lexical entry, e. g. *rapax* 'rapacious; beast of prey' or *togatus* 'wearing a toga; a Roman citizen'. Since LiLa's linking is not contextual, the final decision is to consider these two morphosyntactic categories equivalent for what concerns linking tokens to LiLa;
- misalignments between ADJs and VERBs. This is the case of nominal verb forms considered again as independent lexical entities, the same way as adjectives can be, e. g. *persequens*, so-called present participle of *persequor* 'to follow perseveringly', so 'following perseveringly' or, in a translated sense, 'persecutory'. In LiLa, they are linked as hypolemmas of the respective main verbs.

5 Use case

To show the potentialities of interlinking a subset of texts from the LTA to the other linguistic resources in the LiLa Knowledge Base, a sample query is shown in this section. The query searches for sequences of three lemmas in the CompHistSem texts at hand (see Section 4.1), in the LASLA corpus (Fantoli et al., 2022), in the texts of the 13 books

[32] Univerbated from the phrase *quam ob rem* and opposed to its registration as an adverb in the LiLa lemma bank.

[33] https://universaldependencies.org/u/pos/PART.html

[34] http://www.lemlat3.eu/

[35] In FLL terms, a hypolemma might be seen as an intermediate degree between lemma and word form (cf. Section 3).

[36] https://lila-erc.eu/data/id/lemma/92214

[37] https://lila-erc.eu/data/id/lemma/69949

[38] http://ducange.enc.sorbonne.fr/paterium

of the *Confessiones* by Augustine, taken from *The Latin Library*,[39] in the Index Thomisticus Treebank (IT-TB), which includes texts of Thomas Aquinas (Mambrini et al., 2022), and in UDante, a syntactically annotated corpus featuring the Latin works by Dante Alighieri (Cecchini et al., 2020b). So as to better highlight their characteristics, the works in the LTA's subcorpus are considered separately (splitting parts I-III of the *Decretum Gratiani*) and the LASLA corpus is analyzed per author. This section describes the results of this query limited to token sequences with a frequency of at least 10, up to ten most frequent ones.

Figure 2 shows the text of a SPARQL query. The example in this case is limited to the UDante corpus only for reasons of space. After defining the classes and properties in the relevant ontologies (lines 1-6), the query selects a sequence of three lemmas in the UDante corpus, univocally identified by their URIs (line 11). In order to do that, for every token in the corpus the query selects the next two tokens (lines 8-16) with their respective token labels, their lemmas and lemma labels (lines 17-25). The query then proceeds to order the results by grouping the lemmas by their URIs and puts them in descending order of frequency (lines 26-28). As can be seen from the property `hasLemma` (lines 17, 19 and 21), the LiLa custom ontology provides the linking between a token in the selected corpus and its corresponding lemma in the Lemma Bank, allowing further connections with other lemmatised linguistic resources. This is a pivotal point, as LiLa provides a method to harmonise different lemmatisation criteria, granting interoperability regardless of different citation forms (e. g. *claudeo/claudeor/claudor* 'to limp', all tied to different inflectional paradigms) and/or different written representations (e. g. *sanctus/sancitus* 'saint', originally a participial form of *sancio* 'to establish') of the same lexical item used in specific linguistic resources. [40] The lemma sequences discussed in this section are quoted in small caps and glossed in lowercase translated lemmas, while the examples of textual occurrences are in italics.[41]

The first distinction to be made is that between lemma sequences which are merely grammatical, i. e. sequences composed only of function words such as DE HIC QUI 'from this who' or EX IS QUI 'out-of he who', and sequences with a lexical meaning. The former kind of sequence is quite common among all the works we consider and depends on the language in question, i. e. Latin, and, more in general, on the known Zipfian distribution of words (cf. Newman 2005, §2.1), while the latter is specific to the era and type of each single work.

Considering lexically meaningful sentences, the texts from LTA include sequences which correspond to sentences typical of ecclesiastical language. This is the case with sequences specific to ecclesiastical institutions such as SANCITUS DEUS ECCLESIA 'saint god church', SANCITUS ROMANUS ECCLESIA 'saint roman church': see for example the expressions *sanctae Dei ecclesiae* 'of/to the Holy Church of God', which is also the most frequent sequence of 3 tokens in the *Capitularia Regum Francorum*, and *sanctae Romanae ecclesiae* 'of/to the Holy Roman Church' in the *Decretum Gratiani* I. Other lemma sequences of this kind are ITEM EX CONCILIUM 'also out-of council' and EX CONCILIUM CARTHAGINENSIS 'out-of council carthaginian': see for example *item ex Concilio* 'moreover, from the Council' and *ex Concilio Cartaginensi* 'from the Council of Carthago' which occur in the *Decretum Gratiani* I-IIII. Some other sequences can be considered ecclesiastical insofar as they refer to Christian Latin and lithurgy, such as NOSTER IESUS CHRISTUS 'our jesus christ', IN EXCELSUM DEUS 'in loftiness god', PANIS ET UINUM 'bread and wine', CORPUS ET SANGUIS 'body and blood' and DOMINUS NOSTER IESUS 'lord our jesus': see for example *domini nostri Iesu* 'to our Lord Jesus' in the *Capitularia Regum Francorum*, *in excelsis Deo* 'to God in the highest' in the *De ecclesiasticis officiis* and *panem et uinum* 'bread and wine (accusative case)', *corpus et sanguinem* 'body and blood (accusative case)' and *Dominus noster Iesus* 'our Lord Jesus' in the *Decretum Gratiani* III.

Noting that the most frequently used sequences of tokens in the subset of texts from LTA are *sanc-*

[39] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Confessiones

[40] In the case of different citation forms of the same item belonging to two inflectional categories, e. g. *sequo/sequor* 'to follow' (alternating with respect to morphological active/passive voice), they are considered as two separate lemmas connected via the 'lemma variant' property; if not, e. g. *causa/caussa/kausa/kaussa* 'cause' (all inflecting according to the same nominal paradigm, the so-called "first declension"), they are considered as two written representations of the same lemma; see (Passarotti et al., 2020).

[41] While it is not possible to show all data and tables discussed here for lack of space, they are accessible from a dedicated online repository at https://github.com/CIRCSE/Linking-Computational-Historical-Semantics.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

# get trigram of a corpus
SELECT ?lemmaLabel0 ?lemmaLabel1 ?lemmaLabel2 (count(?corpora) as ?chainCount) ?corpusTitle WHERE {

  VALUES ?corpora {
    <http://lila-erc.eu/data/corpora/UDante/id/corpus>
  }
  ?corpora dc:title ?corpusTitle.
  ?token0 powla:hasLayer/powla:hasDocument/^powla:hasSubDocument ?corpora .
  ?token0 powla:next ?token1.
  ?token1 powla:next ?token2.
  ?token0 lila:hasLemma ?lemmaToken0;
          powla:hasStringValue ?tokenString0.
  ?token1 lila:hasLemma ?lemmaToken1;
          powla:hasStringValue ?tokenString1.
  ?token2 lila:hasLemma ?lemmaToken2;
          powla:hasStringValue ?tokenString2.
  ?lemmaToken0  rdfs:label ?lemmaLabel0.
  ?lemmaToken1  rdfs:label ?lemmaLabel1.
  ?lemmaToken2  rdfs:label ?lemmaLabel2.

} group by ?lemmaToken0 ?lemmaToken1 ?lemmaToken2 ?lemmaLabel0 ?lemmaLabel1 ?lemmaLabel2 ?corpusTitle
order by DESC (?chainCount)
```

Figure 2: Sample query applied to the UDante corpus.

tae Dei ecclesiae and *sanctae Romanae ecclesiae*, one could, if interested in the use of *sanctus* 'saint' in Ecclesiastical Latin, further refine this search with another query to retrieve all the different written representations of the so-called perfect participle of *sancio* 'to establish', of which *sanctus* is a form. In the LiLa Lemma Bank, *sanctus* and its three other possible written representations *sancitus*, *santus* and *xantus* are represented as hypolemmas connected to the lemma *sancio* (cf. Passarotti et al. 2020). In this way, whether in a lemmatised corpus a form like *sanctae* is assigned, for example, the lemma *sancio*, *sancitus* or *sanctus*, in LiLa this lemma is always connected to the same lemma *sancio* and is thus retrievable with a single query. In the specific corpus at hand, this query retrieves 12 participial forms lemmatised under *sancitus*, and 2 785 under *sanctus*: this is a novelty with regards to Classical Latin.

The sequences in the LASLA corpus show a high variety depending on the author. Limiting the data to the sequences of 3 lemmas with frequency greater than 10, the selection includes Caesar, Catullus, Cicero, Seneca and Tacitus. While Caesar is more likely to use strings of lemmas related to spatial descripions and military events such as AD CAESAR MITTO 'to caesar send', SUI IN CASTRA 'self in camp' and EX OMNIS PARS 'out-of all part', the majority of the lemma sequences in Catullus are almost exclusively due to the long and repetitive hymns to Hymenaeus traditionally sung at weddings. Even though the most frequent strings of lemmas in Cicero are mostly due to argumentative purposes (such as UT IS QUI 'as he who' or HAUD SCIO AN 'not know whether'), there are plenty of sequences including typical Republican words such as POPULUS 'people/nation': see for example the sequence POPULUS QUE ROMANUS 'people and roman', which is the only one included in the first 10 most frequent examples, even though other three-lemma sequences such as POPULUS ROMANUS SUM 'people roman be', A POPULUS ROMANUS 'from people roman' and DE PECUNIA REPETO 'from money fetch' refer to institutions and laws of the Roman Republic and have frequency greater than 30.

As for a Christian text like the *Confessiones* by Augustine, even though a generic similarity is due to Christian Latin (see for example the expression DOMINUS DEUS MEUS 'lord god my'), the *Confessiones* are not an ecclesiastical treatise nor a documentary text, but rather a philosophical text based on personal experiences. According to that, its lemma sequences tend to show a peculiar reference to cosmological order (CALEUM ET TERRA 'sky and earth', IN HIC MUNDUS 'in this world') and introspection (IN COR MEUS 'in heart my', IN MEMORIA MEUS 'in memory my').

Thomas Aquinas' *Summa contra gentiles* and the Latin works by Dante Alighieri offer a good example of Medieval Latin from the 13th and 14th centuries. However, the sequences in the *Summa contra gentiles* tend to be due to logic argumentation (SUPRA OSTENDO SUM 'above display be',

UT SUPRA OSTENDO 'as above display', UT OSTENDO SUM 'as display be') according to the rigid exposition of philosophical and theological matters in the Scholastic tradition. The same can be observed in Dante Alighieri's works, where the first 10 lemma sequences are logical sequences useful for speech coherence, as previously observed in Thomas Aquinas' work (ET PER CONSEQUENS 'and for consequence', UT SUPRA DICO 'as above say', PATEO EX PRIMUS 'appear out-of first') except for a broader reference to the universe (CAELUM ET MUNDUS 'sky and world') similar to the CAELUM ET TERRA 'sky and earth' already seen in Augustine and which in Dante is probably a rhetorical device.

These example queries show that the LiLa Knowledge Base makes it possible to extract large quantities of linguistic data (in this case of lexico-textual kind) from several corpora with a single query, covering different eras and genres. This is important when dealing with a language such as Latin, which has a remarkable diachronic and diatopic spread. LiLa also allows for further integrated research with lexical resources such as the *Etymological Dictionary of Latin and the other Italic Languages* (de Vaan, 2008), a valency lexicon (Passarotti et al., 2016), or the prior polarity lexicon of Latin Lemmas *Latin Affectus* (Sprugnoli et al., 2020); see Section 2. In such an interoperable environment, the addition of new resources to the knowledge base allows LiLa to expand its lexical coverage and multiplies the possibilities of connections among (meta)data.

6 Conclusion and Future Work

This paper details the process of linking a subset of the Latin Text Archive, part of the Computational Historical Semantics project, to the LiLa Knowledge Base. This work is part of a wider project which aims to make several linguistic resources for Latin interoperable through LiLa. After years spent building the large collection of lemmas used to interlink distributed resources for Latin, LiLa is now in the phase of exploiting the (meta)data provided by the already available resources to make them interact, assuming that the whole is greater than the sum of its parts.

In such respect, Latin represents a perfect use case where procedures for making linguistic resources interoperable can be developed and tested. Indeed, the history of Latin spans across more than two millennia, showing a wide diversity in terms of genres and provenance of its texts. Moreover, with just a few exceptions, Latin is a dead language, thus making it possible to plan to interlink its entire collection of texts in the (hopefully near) future. Also, the large and diverse community of scholars working on the Latin language, including linguists, philologists, historians and archaeologists, is strictly bound to the empirical evidence provided by Latin texts, as one of the most important sources of information in support of their research work: providing such community with a means to access, query, publish and collect (meta)data from several corpora and lexical resources is a long-time *desideratum* that is finally becoming possible.

In the near future, the *LiLa - Linking Latin* project plans to interlink a number of Latin corpora, including *Musisque Deoque* (Manca et al., 2011), *CRoALa* (Jovanović, 2012), the *Late Latin Charter Treebank* (Korkiakangas, 2021) and the PROIEL treebank (Eckhoff et al., 2018). In the long run, based on the experience of linking a subset of the Computational Historical Semantics corpus, the aim is to link the entire collection of texts provided by the Latin Text Archive to the LiLa Knowledge Base. Given the size and the diversity of the texts therein, this would represent a terrific achievement and advancement for both the communities of Classics and Computational Linguistics.

However, the foundations of LiLa Knowledge Base are built on open and shared formats, models and vocabularies, both to make the resources for Latin interact with each other as well as with those for other languages, and to address the condition of openness that is strictly related to the Linked Data paradigm. Not only are the resources interlinked in LiLa supposed to be openly accessible and downloadable (as the saying goes, "as open as possible, as closed as necessary"), but interlinking the resources is an open process, too. In the Linked Open Data world, everyone is free to add new links between resources: this makes Lila an open-ended knowledge base, which represents the best venue where to publish the digital linguistic resources, in order to set them free from their storage in separate "silos", by making them finally interact. This is the hope of this project: that over the coming years LiLa will grow more and more thanks to the community of developers and providers of linguistic (meta)data for Latin and beyond.

Acknowledgements

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

Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. Ontology Lexicalisation: The *lemon* Perspective. In *9th International Conference on Terminology and Artificial Intelligence (*TIA11*) – Proceedings of the Workshops*, pages 33–36, Paris, France.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 933–942, Marseille, France. European Language Resources Association (ELRA).

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. Association for Computational Linguistics.

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021)*, pages 99–105, Turin, Italy. Associazione italiana di linguistica computazionale (AILC), Accademia University Press.

Christian Chiarcos. 2012a. Interoperability of Corpora and Annotations. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 161–179. Springer, Berlin/Heidelberg, Germany.

Christian Chiarcos. 2012b. POWLA: Modeling Linguistic Corpora in OWL/DL. In *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference,* ESWC *2012, Heraklion, Crete, Greece, May 27–31, 2012, Proceedings*, number 7295 in Lecture Notes in Computer Science, pages 225–239, Berlin/Heidelberg, Germany. Springer.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Roberta Cimino, Tim Geelhaar, and Silke Schwandt. 2015. Digital Approaches to Historical Semantics: New Research Directions at Frankfurt University. *Storicamente*, 11(7).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Number 7 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden, Netherlands; Boston, MA, USA.

Hanne Martine Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Margherita Fantoli, Marco Carlo Passarotti, Eleonora Maria Litta, Paolo Ruffolo, and Giovanni Moretti. 2022. Linking LASLA corpus - LiLa LemmaBank.

Federica Gamba and Daniel Zeman. 2023. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (*UDW, GURT*/SyntaxFest 2023)*, pages 7–16, Washington, D. C., USA. Association for Computational Linguistics (ACL).

James Bradstreet Greenough, George Lyman Kittredge, Albert Andrew Howard, and Benjamin Leonard D'Ooge. 2014. *New Latin Grammar for Schools and Colleges*. Dickinson College Commentaries, Carlisle, PA, USA.

Martin Haspelmath. 2017. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 51(s1000 – Jubilee Issue: 50 Years Folia Linguistica):31–80.

Neven Jovanović. 2012. CroALa. Enhancing a TEI-encoded Text Collection. *Journal of the Text Encoding Initiative*, 2 (Selected Papers from the 2010 TEI Conference).

Timo Korkiakangas. 2021. Late Latin Charter Treebank: contents and annotation. *Corpora*, 16(2):191–203.

Charlton Thomas Lewis and Charles Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford, UK.

Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021. Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021), Milan, Italy, June 29 – July 1, 2022*, Milan, Italy. CEUR-WS.org.

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022. The Index Thomisticus Treebank as Linked Data in the LiLa

Knowledge Base. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 4022–4029, Marseille, France. European Language Resources Association (ELRA).

Massimo Manca, Linda Spinazzè, Paolo Mastandrea, Luigi Tessarolo, and Federico Boschetti. 2011. Musisque Deoque: Text Retrieval on Critical Editions. *Journal for Language Technology and Computational Linguistics*, 26(2 – Annotation of Corpora for Research in the Humanities: *Proceedings of the ACRH Workshop, 5. January 2012, Heidelberg University, Germany*):129–140.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: development and applications. In *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017 conference*, pages 587–597, Leiden, the Netherlands. Lexical Computing CZ s.r.o.

Alexander Mehler, Bernhard Jussen, Tim Geelhaar, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, and Wahed Hemati. 2020. The Frankfurt Latin Lexicon. From Morphological Expansion and Word Embeddings to SemioGraphs. *Studi e Saggi Linguistici*, LVIII(1):121–155.

Mark E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Julianne Nyhan and Marco Passarotti, editors. 2019. *One Origin of Digital Humanities*. Springer Cham, Cham (Zug), Switzerland.

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*, volume 133, pages 24–31, Gothenburg. Linköping University Electronic Press.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, 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*, LVIII(1):177–212.

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'16)*, pages 2599–2606, Portorož, Slovenia. European Language Resources Association (ELRA).

Rachele Sprugnoli, Marco Passarotti, Daniela Corbetta, and Andrea Peverelli. 2020. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3078–3086, Marseille, France. European Language Resources Association (ELRA).

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(160018).