RiCL Research in Corpus Linguistics

Generating linguistically relevant metadata for the *Royal Society Corpus*

Katrin Menzel – Jörg Knappen – Elke Teich
University of Saarland / Germany

Abstract – This paper provides an overview of metadata generation and management for the *Royal Society Corpus* (RSC), aiming to encourage discussion about the specific challenges in building substantial diachronic corpora intended to be used for linguistic and humanistic analysis. We discuss the motivations and goals of building the corpus, describe its composition and present the types of metadata it contains. Specifically, we tackle two challenges: first, integration of original metadata from the data providers (JSTOR and the *Royal Society*); second, derivation of additional linguistically relevant metadata regarding text structure and situational context (register).

Keywords – corpus building and extension; specialized diachronic corpora; written scientific English discourse; *Royal Society Corpus*; register-based metadata

1. INTRODUCTION[1]

This paper provides an overview of metadata generation and technical metadata management solutions for the *Royal Society Corpus* (RSC). The RSC is a diachronic, specialized corpus of scientific English covering more than 330 years of scientific journal articles (1665–1996) with the majority of its texts representing Present-day English and a smaller part representing Late Modern English. The corpus has been built to examine the development of scientific English, that is, the linguistic reaction to specialization and diversification in the scientific domain in terms of style and register/sublanguage formation. Various corpus extensions with more textual and contextual data across several releases have enriched the original corpus version over the years so that the newest releases, RSC 6.0 Open and RSC 6.0 Full (Fischer *et al*. 2020), cover optimized OCR

[1] The work reported in this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 "Information Density and Linguistic Encoding" as well as the Federal Ministry of Education and Research (BMBF) as part of the German Common Language Resources and Technology Infrastructure (CLARIN-D). We are especially indebted to the Royal Society of London and Dr Louisiane Ferlier for making available the source data.

results, more fine-grained linguistic and metadata annotations and a considerably larger number of texts from a much longer time span than previous corpus versions (Kermes *et al*. 2016).

We address two challenges regarding metadata: integration of descriptive metadata from heterogeneous sources and derivation of additional, linguistically relevant metadata from the corpus texts themselves. We first provide an overview of the corpus and the goals and motivation for building it and present the most important metadata requirements (Section 2). We then show which types of metadata have been gathered, distinguishing between descriptive, structural and derived metadata, how they are represented and stored, how they have been checked for completeness, consistency and quality and what types of corrections have been made when deviations were observed (Section 3). Finally, we provide information on the availability of the RSC (Section 4). Section 5 concludes the paper with a brief summary.

2. Overview of the RSC: Corpus material, basic processing and desiderata for metadata

The RSC is a diachronic specialized corpus of scientific English covering more than 330 years of scientific journal articles (1665 to 1996). The primary motivation for building the corpus was to provide a resource for empirically investigating the diachronic development of scientific English (see Halliday and Martin 1993) and its subregisters (sublanguages of chemistry, physics, biology etc.). Another important goal was to create a fairly coherent, homogeneous resource for exploring to what extent the temporal dynamics of language is shaped by communicative concerns, such as efficiency, informativeness, (non-)redundancy and unambiguousness. In particular, we are exploring whether information density (Crocker *et al*. 2016) is an independent factor in language change or whether it correlates with specific extra-linguistic variables, for example, scientific vs. non-scientific domain of discourse (Degaetano-Ortlieb and Teich 2019).

The RSC is embedded in an ecosystem of corpora of English scientific texts such as the *Coruña Corpus of English Scientific Writing* (Moskowich 2012; Moskowich *et al*. 2019), the corpus of *Middle English Medical Texts* (Taavitsainen *et al*. 2005) and its companions for Early and Late Modern English (Taavitsainen and Pahta 2010; Taavitsainen and Hiltunen 2019), or *SciTeX* (Degaetano-Ortlieb *et al*. 2013), a diachronic

corpus of modern English scientific texts. For a discussion and comparison of these corpora to the RSC see Fischer *et al*. (2020).

Going beyond these specific interests, from the beginning, the RSC was built as a resource to be shared by a larger community. As a domain-specific corpus with nearly all full texts from selected prestigious scientific journals that have impacted science across the globe, the RSC is a unique resource for historical linguists and sociolinguists as well as historians of science. As two of the world's longest-running academic journals, the *Philosophical Transactions* and the *Proceedings of the Royal Society of London* used to cover all known scientific disciplines of the time. They split into more specialized series for specific disciplines as the breadth and scope of scientific discovery increased by the end of the nineteenth century to cover mathematical and physical sciences and biological sciences separately. Texts from a few other *Royal Society* journals from the twentieth century, such as *Notes and Records of the Royal Society*, covering the history of science and the history of the *Royal Society* as a scientific community, and the *Biographical Memoirs of Fellows of the Royal Society*, with biographical essays, are also part of the corpus. These can also be queried separately.

The kinds of linguistic studies enabled by the RSC include the diachronic study of selected constructions as pursued, for example, in Construction Grammar, lexical-semantic change, sociolinguistic change, diachronic terminology development and register studies looking at language use according to situational context (field / topic, mode / medium and tenor / attitude of discourse). Metadata on discourse fields, for instance, enable the comparison of different scientific disciplines and help to reveal interesting differences in diachronic developments across disciplines (Teich *et al*. 2016).

2.1. Basic corpus data and processing

The first version of the RSC (2.0) was compiled for the time period of 1665–1869 (ca. 32 million tokens) on the basis of data obtained from JSTOR[2] (Kermes *et al*. 2016) and subsequently enlarged with texts from 1870 to 1996 obtained directly from the *Royal Society* (Fischer *et al*. 2020). We use metadata obtained from the *Royal Society* also for the texts obtained from JSTOR (see Section 3.3). With a size of around 48,000 texts and ca. 300 million tokens, the RSC now contains all English documents of the *Philosophical*

[2] http://www.jstor.org/

Transactions and *Proceedings of the Royal Society of London* and its more specialized successor journals from 1665 to 1996 (see Table 1).

Time period	Tokens
1665–1699	2,582,856
1700–1749	3,414,796
1750–1799	6,342,780
1800–1849	9,112,563
1850–1899	37,313,575
1900–1949	66,051,178
1949–1996	173,147,836

Table 1: *Royal Society Corpus* V5.1.0 (1665–1996)

After OCR optimization, normalization using VARD (Baron and Rayson 2008) was applied and all changes obtained by the normalization procedure were annotated into the corpus. We then added the standard linguistic annotations lemma and part of speech (UPenn tagset) automatically to all of our data using *TreeTagger* (Schmid 1994). In a final step, we added annotations for special research questions, including results of surprisal analysis (Knappen *et al*. 2017). One of the characteristics of electronic corpora is that text elements that are usually of minor importance for linguistic analysis and that are generally difficult to integrate or display correctly in linguistic corpora are typically removed during corpus building. This concerns particularly details of the layout, typographical markup, inserted material such as figures, tables or formulae, which are ignored and removed from the electronic text version. The same applies to elements of the page layout like headings and footers. We also removed hyphenation, even when the hyphenation crosses a page break. However, to also enable studies taking such elements into account, the RSC texts have been linked to their respective source texts on the *Royal Society* journal websites so that visual and layout elements in the image-based PDF files from the scans of the original documents can also be taken into account for individual analyses. The final product is an annotated corpus in the so-called vertical file format (.vrt, see Kermes *et al*. 2016) ready for import into the *Open Corpus Workbench* (Evert and Hardie 2011) and *CQPweb* (Hardie 2012). The .vrt format is a line-oriented file format with one token and all its annotations per line, interspersed by some simple XML-type markup lines. It is not a full XML format because of limitations in tag nesting and because of its line format.

2.2. Requirements on metadata

In accordance with the goals of providing a corpus for linguistic and humanistic study of scientific writing in Late Modern and Present-day English, from the outset, the metadata collected for the RSC provide as much information as possible about potentially relevant extra-linguistic variables. This clearly goes beyond the kinds of 'descriptive metadata' that typically come with datasets provided by digital archives, such as title, author, place etc. Additional metadata need to be derived from the texts themselves or by linking documents up with external sources, such as biographical databases of authors (see also Burnard 2005). Importantly, descriptive metadata and derived metadata have different functions for the user — descriptive metadata are necessary for 'identification' and 'discovery' (e.g. finding a relevant corpus through a data repository), derived metadata enhance the '(re)usability' of a corpus for an intended user community (e.g. facilitating the compilation of subcorpora according to discipline, time period, gender of authors etc.).[3] For the descriptive metadata coming from the text sources, we were faced with the additional challenge of the integration of two sets of metadata; as noted above, our sources came from two different archives (see Section 3.3 below). Two important steps with regard to derived metadata were to mark-up the logical text structure (e.g. title, abstract, text body), henceforth called 'structural metadata', which provides the possibility of integrating text structure elements as factors in analysis, and to assign discourse fields to the documents in the RSC which we realized using topic modelling, henceforth called 'contextual metadata'.

Other desiderata pertaining to formal aspects when a corpus resource is intended for use under FAIR principles are encoding standards (e.g. *Dublin Core*) and technical solutions, such as persistent metadata repositories (see Section 4).

3. RSC METADATA: TYPES OF METADATA, STANDARDS AND TECHNICAL SOLUTIONS

We start by contextualizing the issue of metadata in the context of the FAIR principles of data sharing and show our solutions (Section 3.1). Then we provide an account of the types of metadata we encode, distinguishing between descriptive and derived (structural

[3] See Section 3.1 below for more information on the FAIR principles of data sharing in relation to metadata.

and contextual) metadata (Section 3.2). Finally, we discuss the integration of metadata from heterogeneous sources (Section 3.3).

3.1. Realization of FAIR principles by metadata

The FAIR principles demand that a resource is Findable, Accessible, Interoperable and Reusable (see Table 2). Metadata are necessary for all four FAIR principles. Some of the FAIR principles, namely F4, A1 and A2, also address the necessity of a retrieval infrastructure. This infrastructure is described in Section 4.

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1. the protocol is open, free, and universally implementable
A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be Reusable:
R1. (meta)data are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Table 2: The FAIR Guiding Principles (Wilkinson *et al*. 2016)

The metadata contain a persistent identifier for the corpus in a given corpus version; in our case, a handle from the *Handle System*.[4] The metadata describe the corpus in a rich way, allowing searches for corpora according to a variety of criteria. In this way, the FAIR principles F1–F3 (see Table 2) for Findability are fulfilled. The metadata also contain a description of the corpus, pointers to external resources like publications that describe the corpus and its building process in more detail and information on the copyright of the corpus. These metadata address the FAIR principle R1 (Reusability). The

[4] https://www.dona.net/handle-system

metadata for the whole corpus are provided in two formats, *Dublin Core*[5] and CMDI (Broeder *et al*. 2011). We follow the recommended vocabularies for *Dublin Core*, when applicable. The two formats, *Dublin Core* and CMDI, are standardized and highly interoperable, fulfilling the FAIR principles I1–I3 (Interoperability).

3.2. Types of metadata

3.2.1. Descriptive metadata

In terms of descriptive metadata, each document (text) includes a bibliographical identification of the text in traditional terms (author, journal, volume, pages, year of publication), as well as persistent identifiers to the sources (JSTOR IDs and DOIs from the *Royal Society of London*). This identification again relates to the FAIR principles F1–F3 (Findability) and R1.2 (Reusability) (see Section 3.1). The persistent identifiers enable the users of the corpus to go to photographic scans of the original text directly (see Figure 1 for an example).

Metadata for text *101322*	
Text identification code	101322
Journal in which the article was published	Philosophical Transactions (1665-1678)
Link to the source text on JSTOR	http://www.jstor.org/stable/101322

Figure 1: Excerpt from the metadata view in *CQPweb* showing a direct link to the JSTOR source

Descriptive metadata that provide classificatory information on the texts come from the JSTOR and *Royal Society* data. The *Royal Society* made a choice against relying on software which mines the data to extract titles, authors, dates, etc. and decided to employ indexers to manually catalogue the journals for various data. While we can extract and use these available data to complement our resource, some of them are more important to historians of science than for linguists.

The descriptive metadata are implemented in the .vrt file as attributes to the <text> tag that marks a single text in the corpus. We chose this way of encoding the textual metadata because it is compatible with the intended further processing of the corpus in the *Open Corpus Workbench* (Evert and Hardie 2011). For an overview of all descriptive metadata used in the RSC, see Table 3.

[5] https://dublincore.org/

Metadata type	JSTOR	*Royal Society*
author	✓	✓
title	✓	✓
journal	✓	✓
year	✓	✓
volume	✓	✓
first page	✓	✓
last page	✓	✓
issn	✓	✓
doi		✓
JSTOR id	✓	
language		✓

Table 3: Descriptive metadata taken directly from the sources

3.2.2. Structural metadata

We are concentrating on the text itself and we do not preserve most of its structural layout features, partly because they were not available in our sources (e.g. line breaks are not preserved in parts of the data and paragraphs are not marked), partly because non-linguistic elements like figures, tables or formulae are not directly relevant for linguistic study and are often badly represented in the OCR output. We also do not keep track of typographical markup like italicization. We remove recurring headlines and footers and keep only page breaks in the corpus. Pages are indicated by <page> tags and we add an attribute ID to this tag for the actual page number when we can get at it automatically and reliably. Pages are the only structural units still present in the processed text of the corpus, the titles are available as descriptive metadata to the texts and the abstracts or extracts are available as a separate corpus.

3.2.3. Contextual metadata

To approximate discourse fields, we computed topic models for various versions of the RSC (Fankhauser *et al*. 2016; Bizzoni *et al*. 2020). A topic model is a probability distribution over the words in the texts, and each text is composed from several topics. The topics are learned in an unsupervised fashion, but their labels are assigned manually by inspection of the most salient words. Figure 2 shows the hierarchical clustering of

topics for the RSC 6.0 Open. The five most characteristic words of each topic are given in the Appendix.

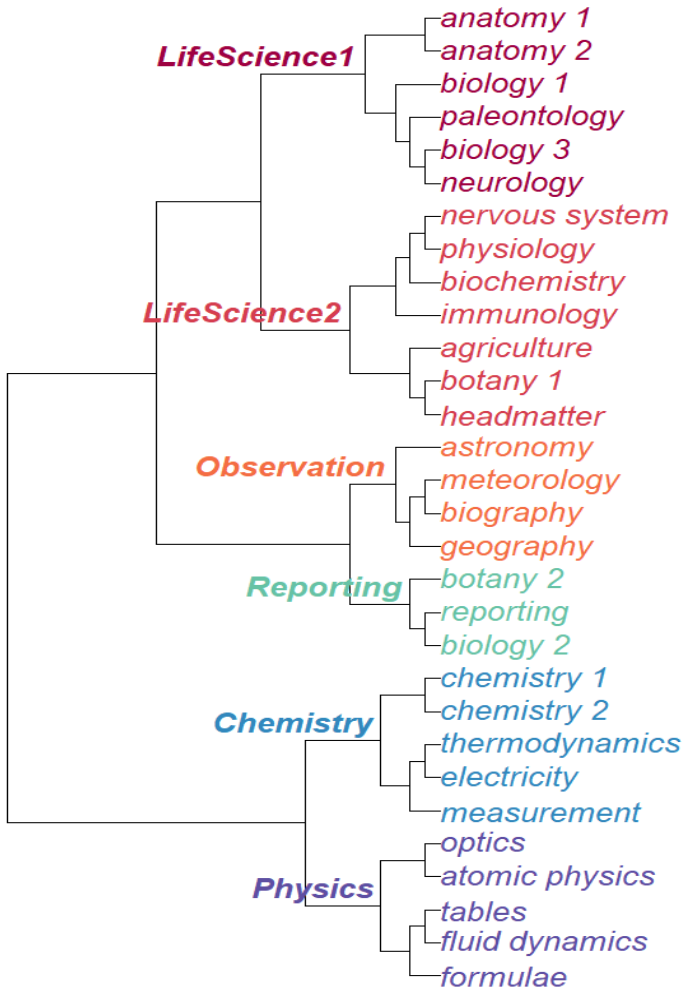

Figure 2: Topic hierarchy derived from topic modelling (RSC 6.0 Open)

As topic models provide not only word-topic but also document-topic assignments, we can add the topic labels as metadata to the documents contained in the corpus, as illustrated in example (1).

(1)
```
<text id="108995" issn="02610523" title="On Hydrofluoric Acid"
[…]
primaryTopic="Chemistry 2"
primaryTopicPercentage="74.1582515464929"
secondaryTopic="Thermodynamics"
secondaryTopicPercentage="12.760468963795098">
```

This is the basis for using topic information as an approximation of the fields of discourse of a text. We encode this information also as a CQP attribute, such that it can be used as a filter in corpus query.

3.3. Integration of metadata

In terms of identification, we set up a match between JSTOR IDs and RS DOIs based on basic bibliographic data: ISSN, volume, year, first page and last page. A match needs to be unique to be considered, as sometimes there are some different items on the same page. We did not use author and title information for this matching, as it decreases the recall significantly due to factors like differences in the encoding of special characters, such as apostrophes or accented letters. Not all articles from the JSTOR sources could be matched to DOIs. Apart from uniqueness there are also different factorings of the material into digital objects, like treating *An accompt on some books* either as single digital object or splitting it into several book reviews, or the treatment of errata and some coding errors. For those articles where DOI and JSTOR ID are matched for texts, the newly obtained metadata from the *Royal Society* are implemented also for previous corpus parts.

The other descriptive metadata types basically match across JSTOR and the *Royal Society* (RS) data. The main difference is that the RS dataset contains some additional and more specific information, such as markup of abstracts or extracts and article titles, contributor information (roles such as author, communicator, biographee or editor; affiliation, e.g. the university name; the *Royal Society* internal identifier number for RS fellows on the basis of which we can also gain further metadata on their biographical data, gender, etc.; election date to the RS), *MathML* markup of mathematical content as well as details on the publishing history.

Integration of the matching types of metadata (see Table 3) was straightforward. For the additional metadata included in the RS bundle, we pursued different strategies. For abstract/extracts (brief summaries of the corpus texts that were either available as abstracts of the respective texts or, in the absence of a given abstract, the first 200 words or the first paragraph of the body of the article), we decided not to add these texts to our corpus metadata but to use this information to create a separate additional corpus that only consists of the abstracts and extracts. Treating the abstracts as a corpus allows us to add linguistic annotations to the abstracts as well. In the case where the RS metadata were

more fine-grained than the ones from JSTOR, we made sure to retain as much detail as possible. For example, for ‘article-type’ for the first 200 years of the corpus we had used the categories ‘full article (fla)’, ‘book review (brv)’, ‘abstract (abs)’, and ‘obituaries (nws)’ in previous corpus versions where ‘fla’, ‘brv’ and ‘nws’ were taken directly from the JSTOR metadata and ‘abs’ was derived from the titles of the articles. For the RSC V6.0 we decided to use the finer grained text types from the *Royal Society* whenever a match was available and to drop the old text types from JSTOR. When no match was found we kept the JSTOR metadata. The vocabulary now includes: abstract, acknowledgement, addendum, appendix, article, astronomical, observation, bibliography, bill of mortality, biography, book review, catalogue, corrigenda, discussion, editorial, errata, experiment, index, lecture, letter, list, magnetical observation, meteorological observation, notes, obituary, preface, report, speech and symposium. Some of the text types like letter, speech or lecture give us a handle on the mode of discourse (e.g. written vs. written-to-be-spoken). For 10,397 texts where we have matched the metadata we see the following correspondence between the text types (Table 4).

We see a good match between the two systems, e.g. ‘brv’ (JSTOR) and ‘book-review’ (RS) are a very good match, and ‘abs’ (JSTOR) corresponds well with ‘abstract’ plus ‘paper-read’ in the RS data. The small category ‘nws’ from JSTOR containing obituary notes on deceased fellows is not represented as a separate article type but absorbed into the category of ‘article’ in the RS data. JSTOR’s ‘fla’ is divided into many subcategories. The majority of the texts, especially the later ones that have a much more standardized format, simply belong to the text type ‘article’. We deleted those texts from the corpus that only consist of tables or other non-text material (e.g. meteorological tables). The RS metadata also have a language attribute with a two-letter ISO 693 code (en, fr, es, la, it, sv, ro). We excluded those articles from the corpus whose main language is not English.

text type	abs	brv	fla	nws
abstract	2,060		560	
appendix			8	
article	35	27	3,421	5
astronomical-observation		1	434	
bill-of-mortality			8	
book-review		227	17	
catalogue			56	
editorial			3	
errata		1	9	
experiment		2	397	
illustration			1	
lecture			63	
letter		3	2,119	
list	1		1	
magnetical-observation			47	
meteorological-observation			134	
notes			3	
paper-read	16		2	
preface			4	
report	4		23	
speech			28	

Table 4: Correspondence between the *Royal Society* text types and our previous text type categories for the first 200 years of the RSC

Metadata concerning the authors of an article and their roles may also be of interest both to linguistic studies and to biographical studies on the authors. In Fischer *et al*. (2018) the authors were annotated and selected manually, matching different spellings of the name of the same person and separating authors with the same name because at that time no further author information was available. For the new release we include the fellowID received from the *Royal Society* whenever available and the author's role in the metadata for each text, as illustrated in example (2).

(2)
```
<text xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:mml="http://www.w3.org/1998/Math/MathML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
id="rsta_1957_0024" issn="0080-4614"

title="The angular acceleration of liquid helium II"
fpage="359" lpage="385"

year="1957" volume="250" journal="Philosophical Transactions
of the Royal Society of London. Series A, Mathematical and
Physical Sciences"

author="H. E. Hall|D. Shoenberg, F. R. S."
fellowID="NA4060|NA5281"
```

```
authorRole="author|communicator" type="article"
corpusBuild="6.0"
doiLink="http://dx.doi.org/10.1098/rsta.1957.0024"
language="en">
```

From the *Royal Society* metadata on authors we use the fellowID (e.g. NA8137, named 'Code' in Table 5) uniquely identifying a fellow of the *Royal Society* and the authorRole when available. With the fellowID more biographical data of that specific author can be obtained. We have not added this additional information to the corpus yet as it needs some additional processing, but the fellowID is sufficient to link up to the information when needed. In the future, we intend to add nationality, gender (female first names of text authors or co-authors are often either spelled out or accompanied by the information *Miss*/*Mrs* in front of the initials of the first names) and the author's age (to be calculated from the birth date and year of publication if available).

Fellow details	
Surname	Boyle
Forenames	Robert
Epithet	Natural Philosopher and Chemist
Dates of Existence	1627 - 1691
Nationality	British
Dates and Places	Birth: Lismore Castle, Munster, Ireland (25 January 1627)
Address	Stalbridge Manor, Dorset (1645-1655); Oxford (1655-1668) Lady Ranelagh's house, Pall Mall, London (1668-1691)
Activity	Research Field: Natural philosophy, physics, chemistry
RS Activity	Membership: Founder Fellow Election Date: 28/11/1660 Council: Elected and declined Presidency of the Royal Society (1680)
Relationships	Fourteenth child, seventh son of Richard Boyle, 1st Earl of Cork, and his second wife, Catherine, daughter of Sir Geoffrey Fenton, Principal Secretary of State for Ireland [...]
Code	NA8137

Table 5: Example of Fellow details from the Royal Society Fellows Directory[6]

The author role helps us to identify who has actually written the article, who has communicated it, or who was taking part in a different role, for example, as an author of a reviewed book or as a biographee. This is useful to select works actually written by a

[6] https://royalsociety.org/fellows/fellows-directory/

certain author, for example, in order to determine the author's style or the development of an author over time. Many texts in the Late Modern English part of the RSC were submitted either by single individual authors who were Fellows of the Royal Society or by pairs of individual non-members and Fellows where the latter typically only acted as 'communicators'. Some prominent Fellows steered a large number of papers by non-Fellows through the publication process, often without having contributed to the actual research (cf. also Harrison 1989: 112). The proportion of multi-author papers has generally increased over time. Articles written by research teams become a common form in the Present-day English part of the RSC where it is not unusual to find research articles with four to ten authors, co-authors and other discourse participants.

4. Availability of the RSC

The corpus is deposited at a data repository at the certified CLARIN center of Saarland University.[7] CLARIN centers offer both direct web access to the metadata and an OAI-PMH interface for metadata harvesting. This guarantees that the corpus metadata are publicly accessible, addressing the FAIR principles A1 and A2 (accessibility). Large parts of the RSC have already been made available for free download and online query in a *CQPweb* interface from the CLARIN-D center at Saarland University under a persistent identifier.[8] Compared to the current release (V4.0), the next open version (V6.0 Open; Fischer *et al*. 2020) covers 50 additional years. Texts from certain decades currently remaining under copyright are not available for download as full texts, but the full version is available onsite. The CLARIN *Virtual Language Observatory* (VLO) harvests the metadata of the corpus and provides a facet search for corpora and language resources. The various elements in the CMDI metadata are mapped to the facets of the VLO and can be used to restrict search results (Van Uytvanck *et al*. 2012). This makes the RSC visible and fulfils the FAIR criterion F4.

5. Conclusion

We have shown how metadata contribute to the fulfilment of the FAIR principles and add value to a corpus for re-use by other researchers. We also note that metadata alone are not

[7] https://www.clarin.eu/content/clarin-centres
[8] http://hdl.handle.net/11858/00-246C-0000-0023-8D1C-0

enough to fulfil all FAIR principles: a retrieval infrastructure is also required. We used the *Royal Society Corpus* as a relevant example of how to obtain metadata, how to integrate them from different sources, and how to add some contextual metadata using topic modelling. For a summary of the metadata we discussed in this article see Table 6.

author	descriptive	doi	descriptive
first page	descriptive	last page	descriptive
title	descriptive	journal	descriptive
year	descriptive	volume	descriptive
issn	descriptive	JSTOR id	descriptive
language	descriptive	page	structural
primary topic	contextual	primary topic percentage	contextual
secondary topic	contextual	secondary topic percentage	contextual

Table 6: (Types of) metadata discussed in this article

The metadata provided for the RSC 6.0 Open allow for differentiated corpus analysis and query according to linguistically relevant variables such as time, author and topic (field of discourse) by selecting a subcorpus or comparing two or more subcorpora according to the metadata.

REFERENCES

Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics.* Birmingham, UK: Aston University. http://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Peter Fankhauser and Elke Teich. 2020. Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence – Language and computation, Research topic Computational Sociolinguistics* 3, Article 73.

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck and Andreas Witt. 2011. A pragmatic approach to XML interoperability – The Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies 7.*

Burnard, Lou. 2005. Metadata for corpus work. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books, 30–46.

Crocker, Matthew W., Vera Demberg and Elke Teich. 2016. Information density and linguistic encoding (IDeaL). *Künstliche Intelligenz* 30: 77–81.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ekaterina Lapshinova-Koltunski and Elke Teich. 2013. SciTex: A diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, Martin Durrell, Silke Scheible and Richard J. Whitt eds. *New Methods in Historical Corpora. Volume 3 of Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP).* Tübingen: Narr, 93–104.

Degaetano-Ortlieb, Stefania and Elke Teich. 2019[online]. Towards an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0088

Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Paper 153. Birmingham, UK: University of Birmingham. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf

Fankhauser, Peter, Jörg Knappen and Elke Teich. 2016. Topical diversification over time in the *Royal Society Corpus*. In Maciej Eder and Jan Rybicki eds. *Digital Humanities 2016: Conference Abstracts*. Kraków, Poland: Alliance of Digital Humanities Organizations (ADHO), 496–500. https://dh2016.adho.org/abstracts/322

Fischer, Stefan, Jörg Knappen and Elke Teich. 2018. Using topic modelling to explore authors' research fields in a corpus of historical scientific English. In *Digital Humanities 2018: Book of Abstracts*. Mexico City, Mexico: Alliance of Digital Humanities Organizations (ADHO), 581–584. https://dh2018.adho.org/en/using-topic-modelling-to-explore-authors-research-fields-in-a-corpus-of-historical-scientific-english/

Fischer, Stefan, Jörg Knappen, Katrin Menzel and Elke Teich. 2020. The *Royal Society Corpus* 6.0: Providing 300+ years of scientific writing for humanistic study. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blace, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 794–802. https://www.aclweb.org/anthology/2020.lrec-1.99.pdf

Halliday, Michael A.K. and James R. Martin eds. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer.

Hardie, Andrew. 2012. *CQPweb* – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17: 380–409.

Harrison, Andrew John. 1989. *Scientific Naturalists and the Government of the Royal Society 1850–1900.* The Open University, PhD dissertation.

Kermes, Hannah, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen and Elke Teich. 2016. The *Royal Society Corpus*: From uncharted data to corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Sterlios Piperidis eds. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association, 1928–1931. https://www.aclweb.org/anthology/L16-1305.pdf

Knappen, Jörg, Stefan Fischer, Hannah Kermes, Elke Teich and Peter Fankhauser. 2017. The making of the *Royal Society Corpus*. In Gerolf Bouma and Yvonne Adesam eds. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Gothenburg, Sweden: Linköping University Electronic Press, 7–11. https://www.aclweb.org/anthology/W17-0503.pdf

Moskowich, Isabel. 2012. CETA as a tool for the study of modern astronomy in English. In Isabel Moskowich and Begoña Crespo eds. *Astronomy "Playne and Simple": The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins, 35–56.

Moskowich, Isabel, Begoña Crespo, Luis Puente-Castelo and Leida Maria Monaco eds. 2019. *Writing History in Late Modern English – Explorations of the Coruña Corpus.* Amsterdam: John Benjamins.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Taavitsainen, Irma, Päivi Pahta and Martti Mäkinen eds. 2005. *Middle English Medical Texts*. Amsterdam: John Benjamins.

Taavitsainen, Irma and Päivi Pahta eds. 2010. *Early Modern English Medical Texts: Corpus Description and Studies*. Amsterdam: John Benjamins.

Taavitsainen, Irma and Turo Hiltunen eds. 2019. *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*. Amsterdam: John Benjamins.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes and Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data mining approach using register features. *Journal of the Association for Information Science and Technology (JASIST)* 67/7: 1668–1678.

Van Uytvanck, Dieter, Herman Stehouwer and Lari Lampen. 2012. Semantic metadata mapping in practice: The Virtual Language Observatory. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, Stelios Piperidis eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).* Istanbul, Turkey: European Language Resources Association, 1029–1034. http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf

Wilkinson, Mark D., Michel Dumontier, [...] and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Corresponding author
Katrin Menzel
Department of Language Science and Technology
Saarland University, Saarbrücken Campus
66123 Saarbrücken
Germany
e-mail: k.menzel@mx.uni-saarland.de

received: February 2020
accepted: October 2020

APPENDIX

The following table gives the most characteristic words (word forms) for each of the thirty topics from the topic model in Section 3.2.3.

anatomy 1	fig plate cartilage part skull
anatomy 2	fig bone bones teeth surface
biology 1	number eggs species larvae female
paleontology	fig plate species form structure
biology 3	cells fig cell tissue nucleus
neurology	fibres posterior anterior fig side
nervous system	nerve muscle contraction stimulation muscles
physiology	blood serum action normal pressure
biochemistry	solution cent water acid vol
immunology	days growth water found bacteria
agriculture	nitrogen soil plants plot years
botany 1	fig plate section cells plants
headmatter	vol society london des der
astronomy	sun observations time stars distance
meteorology	observations days day p.m. magnetic
biography	society work years royal professor
geography	feet water sea found miles
botany 2	leaves plants plant fig species
reporting	great time made found account
biology 2	animal part blood parts body
chemistry 1	water air experiments quantity heat
chemistry 2	acid solution water obtained salt
thermodynamics	temperature pressure air gas tube
electricity	current wire resistance magnetic positive
measurement	inch fig inches made length
optics	light rays glass colour red
atomic physics	lines spectrum line bands spectra
tables	values table curve results case
fluid dynamics	velocity surface motion force direction
formulae	equation equations function form cos