

Making Sign Language Research Findable: The sign-lang@LREC Anthology and the Sign Language Dataset Compendium

Marc Schulder, Thomas Hanke, Maria Kopf

Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany

{marc.schulder, thomas.hanke, maria.kopf}@uni-hamburg.de

Abstract

Resources and research on sign languages are sparse and can often be difficult to locate. Few centralised sources of information exist. This article presents two repositories that aim to improve the findability of such information through the implementation of open science best practices. The *sign-lang@LREC Anthology* is a repository of publications on sign languages in the series of sign-lang@LREC workshops and related events, enhanced with indices cataloguing what datasets, tools, languages and projects are addressed by these publications. The *Sign Language Dataset Compendium* provides an overview of existing linguistic corpora, lexical resources and data collection tasks. We describe the evolution of these repositories, covering topics such as supplementary information structures, rich metadata, interoperability, and dealing with the challenges of reference rot.

1 Introduction

Sign language linguistics is both a small and young field, compared to research on spoken languages. This is especially true for areas such as computational and corpus sign linguistics, which only became feasible with the advent of high-quality digital media, as signs require video imagery to be represented suitably. In recent decades, these areas of research have grown markedly, as has the number of digital sign language resources, such as corpora and lexica. Nevertheless, data availability for individual sign languages continues to range from sparse to virtually non-existent (Morgan et al., 2022). Finding these precious resources or the research relating to them can often require extensive web searches or literature review in several languages, as few centralised sources of information exist.

In this article we present two repositories we created to support sign language researchers in their work by compiling metadata-rich collections of sign language research articles and datasets.

The *sign-lang@LREC Anthology*¹ is the open-access publication repository of the *Workshop Series on the Representation and Processing of Sign Languages* (see fig. 1). To date the Anthology covers 485 articles: 370 sign-lang workshop papers and an additional 115 papers from co-located events. Apart from bibliographic metadata, each article is enhanced with information on the languages, datasets, tools, and project affiliations central to it, allowing researchers a more focussed search for relevant literature. While the Anthology was released in 2020, this is the first article to describe it.

The *Sign Language Dataset Compendium*² is a structured overview of existing linguistic resources on sign languages (see fig. 2). It covers 43 corpora and 86 lexical resources across 82 sign languages, as well as 28 data collection tasks commonly used in the described corpora. Since its introduction in Kopf et al. (2022a) it has received several updates. Apart from the addition of 25 new entries (including resources for 10 more sign languages) and the maintenance of existing materials, various features were added and improved, which we will describe in this article.

Both repositories embrace FAIR principles (Wilkinson et al., 2016) by exposing rich metadata about themselves and the resources they document, building on open standards and providing stable identifiers wherever possible. At the same time they have to deal with the challenges of reference rot (Klein et al., 2014) as external references change, move, and disappear.

The article is structured as follows: Section 2 provides relevant background information on sign language research (section 2.1) and existing repositories (section 2.2). Section 3 provides general introductions to the sign-lang@LREC Anthology (section 3.1) and Sign Language Dataset Compendium

¹<https://doi.org/10.25592/dgs.lrec>

²<https://doi.org/10.25592/dgs.sldc>

sign-lang@LREC Anthology

[Proceedings](#) | [Authors](#) | [Projects](#) | [Languages](#) | [Data](#) | [Tools](#) | [Network](#)

Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of Greek Sign Language and German Sign Language

[Bigeard, Sam](#) | [Schulder, Marc](#) | [Kopf, Maria](#) | [Hanke, Thomas](#) | [Vasilaki, Kyriaki](#) | [Vacalopoulos, Anna](#) | [Goulas, Theodoros](#) | [Dimou, Athanasia-Lida](#) | [Fotinea, Stavroula-Evita](#) | [Efthimiou, Eleni](#)

Volume: Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources

Venue: Marseille, France

Date: 26 June 2022

Pages: 9–15

Publisher: European Language Resources Association (ELRA)

License: CC BY-NC 4.0

sign-lang ID: 22036

ACL ID: 2022.signlang-1.2

ISBN: 978-10-95546-86-3

Content Categories

Projects: DGS-Korpus project, EASIER

Languages: German Sign Language, Greek Sign Language, German, Greek

Corpora: DGS Corpus, POLYTRIPON Parallel Corpus

Dictionaries: NOEMA

Lexical Databases: GermaNet, Greek WordNet, Multilingual Sign Languages Wordnet

Abstract

Wordnets have been a popular lexical resource type for many years. Their sense-based representation of lexical items and numerous relation structures have been adopted for a variety of computational and linguistic applications. The inclusion of different wordnets into multilingual wordnet networks has further extended their use into the realm of cross-lingual research. Wordnets have been released for many spoken languages. Research has also been carried out into the creation of wordnets for several sign languages, but none have yet resulted in publicly available datasets. This article presents our own efforts towards an inclusion of sign languages in a multilingual wordnet, starting with Greek Sign Language (GSL) and German Sign Language (DGS). Based on differences in available language resources between GSL and DGS, we trial two workflows with different coverage priorities. We also explore how synergies between both workflows can be leveraged and how future work on additional sign languages could profit from building on existing sign language wordnet data. The results of our work are made publicly available.

Document Download

[Paper PDF](#) | [Poster](#) | [BibTeX File](#) | [Abstract](#)

Cite as

Citation in ACL Citation Format

Sam Bigeard, Marc Schulder, Maria Kopf, Thomas Hanke, Kyriaki Vasilaki, Anna Vacalopoulos, Theodoros Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, Eleni Efthimiou. 2022. *Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of Greek Sign Language and German Sign Language*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 9–15, Marseille, France. European Language Resources Association (ELRA).

BibTeX Export

```

@inproceedings{bigeard:22036:sign-lang:lrec,
  author = {Bigeard, Sam and Schulder, Marc and Kopf, Maria and Hanke, Thomas and Vasilaki, Kyriaki and Vacalopoulos, Anna and Goulas, Theodoros and Dimou, Athanasia-Lida and Fotinea, Stavroula-Evita and Efthimiou, Eleni},
  title = {Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of (Greek) (Sign) (Language) and (German) (Sign) (Language)},

```

Figure 1: sign-lang@LREC Anthology article entry. Shown are top menu, title, authors, publication details, identifiers (including ACL Anthology link), content categories, abstract, downloads for paper, BibTeX and supplementary materials, and citation recommendations in text and BibTeX format.

(section 3.2). Section 4 highlights various aspects of interoperability, such as the rich metadata provided by the repositories (section 4.1), how we connect to different resources (section 4.2), our shared inventory of language metadata (section 4.3), and other synergies in workflows and data structures (section 4.4). In section 5 we address the threat of reference rot by leveraging semi-automated availability checks, archival copies and robust links. In the remaining sections we provide discussions of the impact that our repositories have had to date (section 6), their limitations (section 7) and offer concluding words (section 8).

2 Background

2.1 Challenges in Sign Language Research

Working on and with sign language resources and technologies involves a number of challenges resulting from both the specific requirements of sign languages and the relative youth of the field.

Sign languages have no commonly used written forms, so textual annotation often relies on glossing,

The Sign Language Dataset Compendium

[Start](#) | [About](#) | [Corpora](#) | [Lexical Resources](#) | [Tasks](#) | [Languages](#) | [Credit](#) | [More](#)

Corpus

ECHO Corpus

The European Cultural Heritage Online (ECHO) corpus is a multilingual corpus containing video material from three SLs: [NLS](#), [BSL](#), and [STS](#). Eight signers were recorded for 1.5 hours following the same tasks in each language. For [NET](#) and [BSL](#), sign language poetry was added to the corpus. Additionally annotated segments of the GeHörlos Sol corpus of [DGS](#) ([Heßmann, 2001](#)) were added to the corpus. The ECHO project was a 18-month EU funded project dedicated to bring Essential Cultural Heritage online. The ECHO corpus was built from 2005–2008 by the Max Planck Institute for Psycholinguistics, Radboud University and University of Lund.

Filming took place in a studio with one or two signers at the same time. The signers were sitting or standing and depending on the task, recorded separately or closely next to each other. A single-coloured background was used.

Languages	British Sign Language , Sign Language of the Netherlands , Swedish Sign Language , German Sign Language
Size	1.5 hours recorded
Participants	8 participants Native signers 20–40 years old
Metadata Format	IMDI, CLAC
Translation	Dutch, English and Swedish, size unknown
Annotation	See Nonhebel et al. (2004)
Data Format	ELAN
License	CC BY-NC-ND 3.0
Access	Open access to videos and transcripts via Language Archive
Webpages	Project page: http://echo2.mpiwg-berlin.mpg.de/home/ (archival copy) Project results: http://www.lit.ru.nl/sign-lang/echo/ (archival copy) Dataset: https://hdl.handle.net/1839/00-0000-0000-0001-4892-C
Institution	Max Planck Institute for Psycholinguistics, Radboud University Nijmegen, University of Lund

Cite as

Barbara Cassin, Wim Emmerik, Annika Nonhebel, Els van der Kooij, Johanna Mesch, Annemieke van Kampen, Onno Crasborn, Rachel Sutton-Spence, Rachel Sutton-Spence / Dafydd Waters, Anja Hiddinga, British Broadcasting Corporation (BBC), Dafydd Waters, and Leendert Pot. (2003–2005). Collection "ECHO": The Language Archive. <https://hdl.handle.net/1839/00-0000-0000-0001-4892-C>. (Accessed [insert date])

Common tasks used in this corpus

• Hide/Show tasks

Task	Lexical elicitation
Corpus Language	British Sign Language
# recordings – open access	1
# recordings – restricted access	0
Data available	https://hdl.handle.net/1839/00-0000-0000-0001-49AF-B

Task	Lexical elicitation
Corpus Language	Sign Language of the Netherlands
# recordings – open access	4
# recordings – restricted access	0
Data available	https://hdl.handle.net/1839/00-0000-0000-0001-4A68-D

Figure 2: Example of a corpus entry in the Compendium. Shown are the top menu, free-form description, info table, citation recommendation and start of the list of data collection tasks. Not shown is the list of references and links to other information sources.

i.e. representing a sign through a rough lemma-level translation to a written language. This introduces various complications, such as ensuring a sign is always annotated with the same gloss, distinguishing synonymous but distinct signs that may be glossed using the same translation, encoding morphosyntactic information, and annotating multiple simultaneous language channels (two hands and various non-manual components). While a baseline consensus for glossing conventions grew from the Auslan Corpus annotation guidelines ([Johnston, 2007](#)), annotation practices still vary heavily across corpora ([Kopf et al., 2022b](#)), making it difficult to compare or combine resources ([Schulder et al., 2023](#); [De Sisto et al., 2022](#)).

While corpora rely on the vocabulary of lexical resources to ensure consistent annotation, lexica in turn depend on corpora as a source of discovery of that vocabulary and its actual use. Creating either resource is a costly endeavour: preparing an hour of data can easily take 60 hours of work for basic annotation ([Hochgesang et al., 2023](#)) and up to 1000 hours for full publication ([Schulder and Hanke, 2022](#)). NLP pipelines to support resource

creation, so ubiquitous for many spoken languages, do not yet exist for sign languages. In a classic catch-22, sign language NLP research is hindered by the extreme sparsity of annotated sign language data which it seeks to remedy. Combining datasets, possibly across languages, is one possible way to alleviate this issue, but it requires researchers to find suitable datasets that can be harmonised not only regarding primary video materials, but also in terms of annotation (Morgan et al., 2022).

2.2 Repositories

2.2.1 Publication Repositories

Among repositories for academic publications, the one most impactful and relevant to our work has been the ACL Anthology³ (Bollmann et al., 2023; Gildea et al., 2018). Operated by the *Association of Computational Linguistics (ACL)*, it is a large repository of over 100.000 open access publications in the field of computational linguistics and related areas. Its code base and publication metadata are both open source and its development driven in large parts by community volunteers. It covers all publications by the ACL, as well as those of numerous other venues and organisations, including the majority of LREC conference proceedings. While originally only the LREC main conferences were covered, from LREC 2020 onwards it also includes the proceedings of the LREC satellite workshops, including those of sign-lang@LREC.

The ACL Anthology is a strong example of an open data repository and we have taken inspiration from many of its features, such as citation export formats, Zotero integration and the pivot to static HTML pages and metadata formats described in Bollmann et al. (2023).

2.2.2 (Meta)Data Repositories & Surveys

Information on sign language datasets can be found in a number of repositories. These may be archives of the data itself which expose metadata for their content or metadata repositories that reference external sources of data.

Hosting sign language datasets is a non-trivial task. Given the size of high resolution video recordings and the best practice of simultaneously recording sign language data from multiple angles (Hanke et al., 2010), the storage demands for corpora are terabytes for legacy SD video (Johnston and Schembri, 2006), hundreds of terabytes for HD video (DGS-

Korpus, 2022) and will reach petabytes as the field moves towards 4K and 6K resolutions as new standards. These demands usually have to be addressed by the institution at which the resource was created, but may also be deposited with a suitable data archive.

Among the datasets we document, two archives stood out for the number of corpora they contain and their support for metadata specific to sign languages: *The Language Archive*⁴, hosted by the Max Planck Institute for Psycholinguistics in Nijmegen, and the *Endangered Languages Archive*⁵, run by the Berlin-Brandenburg Academy of Sciences and Humanities. Together they account for the storage of twenty of the datasets documented in the Compendium.

Given the distribution of datasets across many institutions, another way to centralise information and make data more findable are metadata repositories. Among the repositories for language data that also contain entries on sign language datasets are the Open Language Archives Community⁶ (OLAC) (Simons and Bird, 2003), the CLARIN Virtual Language Observatory⁷ (VLO) (Van Uytvanck et al., 2012; Goosen and Eckart, 2014), Meta-Share⁸ (Federmann et al., 2012), the European Language Grid⁹ (ELG) (Rehm et al., 2021) and the LRE Map¹⁰ (Calzolari et al., 2010).

These repositories mainly build on collecting information from numerous sources through metadata harvesters. Inclusion in this syndication process may require an application process (OLAC) or be mostly focussed on member institutions of a network (VLO, Meta-Share). They may even build on collating information from other (meta)data repositories, as is the case for ELG. The one exception to this approach is the LRE Map, which relies on resource creators submitting information directly, primarily as part of the article submission process for LREC conferences.

An entirely different type of information source are surveys and curated resource tables like, for example, Schmaling (2012), Konrad (2012), Moryossef and Goldberg (2021) or the CLARIN Resource family page for sign language resources¹¹.

⁴<https://archive.mpi.nl/tla/>

⁵<https://www.elararchive.org/>

⁶<http://www.language-archives.org/>

⁷<https://vlo.clarin.eu/>

⁸<http://metashare.ilsp.gr/>

⁹<https://live.european-language-grid.eu>

¹⁰<https://lremap.elra.info>

¹¹<https://www.clarin.eu/resource-families/sign-language-resources>

³<https://aclanthology.org/>

The Sign Language Dataset Compendium presented in our article falls between these resource types, combining regular updates with the editorial practices of a survey and the rich metadata of a repository. Since its latest release, the Compendium also includes a section on further sources of information, listing the aforementioned repositories and surveys as well as additional ones.

3 The Repositories

3.1 The sign-lang@LREC Anthology

The *Workshop Series on the Representation and Processing of Sign Languages (sign-lang@LREC)* was started in 2004 as a satellite event of the *International Conference on Language Resources and Evaluation (LREC)* and has been a part of every LREC conference since.¹² It provides a forum for work on sign language resources and technologies, bringing together researchers from a variety of fields, such as linguistics, natural language processing and computer vision.

As with other LREC workshops, the sign-lang@LREC proceedings are published by ELRA and made available through the website of that year's conference. As is common practice, each year's workshop also has its own website to communicate information, such as its call for papers and the workshop programme. It also offers authors the option to publish supplementary materials like signed video presentations and PDFs of posters or slide sets. As an additional service to conference attendees, each workshop website also lists all main conference presentations related to sign languages.

3.1.1 Creating the Anthology

In 2020, we introduced the *sign-lang@LREC Anthology* to create a central location for publications of the entire workshop series. While the focus of the workshop websites lies on communicating information before and during their respective event, the Anthology would be the post-event repository of workshop outputs. Following the traditions of the workshop websites, the Anthology covers not only publications of the workshop, but also sign language papers from the LREC main conference and its other workshops, and gives authors the option to provide supplementary presentation materials.

Half a year after the release of the sign-lang@LREC Anthology, the inclusion of LREC

¹²The first two authors of this article are members of the sign-lang@LREC organising committee.

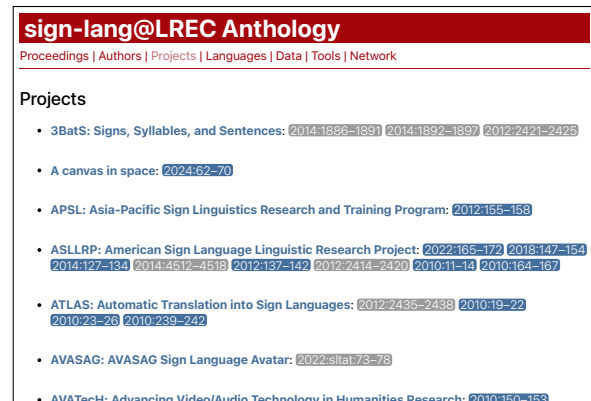


Figure 3: Top of the Anthology project index. To date, the index covers about a hundred projects. Each is shown with its name, followed by a list of its publications. Each publication is given using its Anthology ID and a hover tooltip showing its reference entry. Blue boxes indicate sign-lang@LREC workshop papers, grey boxes indicate papers from other events. Each project name and paper ID is a link leading to its respective entry.

workshops in the ACL Anthology became known, raising the question whether maintaining a separate repository was sensible. As the ACL Anthology ingestion did not include workshops of previous years, we decided to continue our efforts and to look for ways to enrich our repository that were tailored to the needs of our community, such as the categorial indices discussed in the following section.

3.1.2 Categorial Indices

Articles in the Anthology can be accessed through a number of different indices, allowing users different perspectives through which to look for publications. In addition to the usual groupings by **proceedings** or **author**, papers may also be grouped by **languages** that they address, the **datasets** and **tools** that they introduce or make use of, and the **projects** that they originate from (see fig. 3). Language, data and tool indices are sub-grouped further, e.g. separating signed, spoken and tactile sign languages or corpora, dictionaries and other lexical resources.

The indices allow users increased flexibility in tailoring their search to their own needs, e.g. by focussing on a specific language, comparing different tools or compiling the outputs of a specific project to identify resources with high compatibility.

Each index entry has its own page in which it lists its publications, just as each publication page specifies all its index entries. Index entries also provide additional information, specific to their category. Author profiles specify ORCID IDs (Haak et al., 2012), while projects, datasets and tools pro-

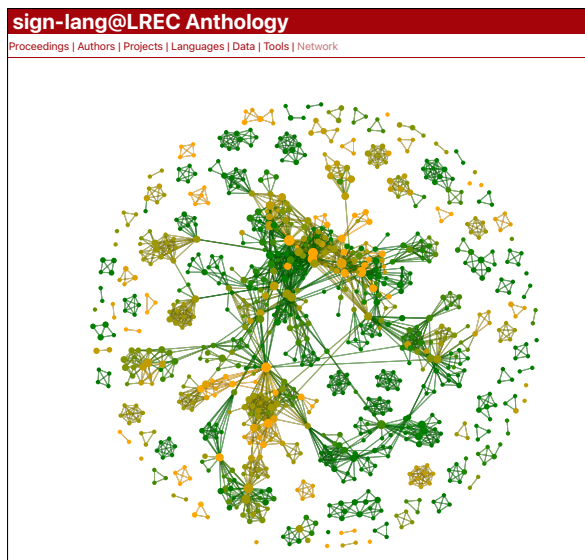


Figure 4: The Anthology author network graph. Each dot represents one author and each line the co-authorship between two authors. The bigger a dot or thicker a line, the more (co-)publications are present. Hovering over a dot shows the name of the author and clicking on it leads to the author’s profile page.

vide relevant URLs and common name variations (acronym, short form, local and English name), plus license information in the case of datasets and tools. There are also links between categorial entries, connecting closely related datasets with each other (e.g. a co-created corpus and lexicon pair) as well as linking projects to the datasets and tools that were produced through them. Languages specify their ISO 639-3 and Glottolog codes, names and acronyms (see section 4.3 for details).

The interconnectedness of the sign language research community is also highlighted in a network graph that visualises co-authorship patterns across all Anthology publications, showing research group clusters and how they collaborate (see fig. 4).

3.1.3 Citation and Export Formats

Like the ACL Anthology and other publication repositories, the sign-lang@LREC Anthology makes article metadata available in various formats. BibTeX reference entries can be downloaded for specific articles, whole proceedings or the entire Anthology. Individual articles also provide a reference entry text for easy copy-pasting, preformatted in ACL reference style. Embedded metadata also allows direct import of publications into reference managers (see section 4.1 for details).

3.2 The Sign Language Dataset Compendium

The *Sign Language Dataset Compendium* provides a curated overview of existing linguistic resources on sign languages, with free-form descriptions of each resource as well as structured information regarding common aspects like dataset size, languages covered, usage licence, file and metadata formats and relevant URLs. It covers linguistic corpora and lexical resources, as well as an inventory of commonly used data collection tasks, cross-matched to the corpora that contain them.

The Compendium originated with Kopf et al. (2021), a report for the EASIER project¹³ in which we provided an overview of existing resources for European sign languages. The report in turn used a comprehensive review of the sign-lang@LREC Anthology as a main source of information, supplemented by further web and literature reviews and personal communications with resource creators. This same review was the basis of the first version of the categorial indices (see section 3.1.2). Following strongly positive responses to the report, we expanded it further into the first release of the Compendium (Kopf et al., 2022a), growing its scope to global coverage of sign languages and making it available both as a website and as a static document.

The Compendium continues to receive updates whenever we encounter new eligible resources in the course of our ongoing work in the domain of resource-driven sign language linguistics. Resource creators and other researchers have also begun to explicitly contact us to make us aware of new resources as well as to provide additional information for entries.

Kopf et al. (2022a) provides a detailed discussion of the curation criteria of the Compendium and of the information categories provided in each entry. Since then, a series of refinements have been applied to the structure of entries: A “*Cite as*” field was added to specify the creators’ recommended way to cite their dataset. The reference list for individual entries now differentiates between articles on the resource itself and other works cited for context. Other improvements will be described in the following sections, such as the production of rich metadata (section 4.1), referencing and connecting with other resources (section 4.2), information shared across resources (sections 4.3 and 4.4), and dealing with reference rot (section 5).

¹³See <https://www.project-easier.eu/> as well as <https://doi.org/10.3030/101016982>

4 Interoperability

The goals of both the Anthology and Compendium are to make information more easily findable and provide a net benefit to the research community. To this end we build on establishing connections at various levels, be it between entries or between repositories, by referencing external sources, exposing our information for processing, or by sharing data and code across resources.

4.1 Metadata

Following FAIR principles, our repositories provide rich metadata that describes the repositories themselves and the resources which they document. To support a number of different use cases, we serve metadata in a variety of schemas.

Most metadata is served through the head section of individual HTML pages. General metadata is served using Dublin Core¹⁴. Open Graph¹⁵ helps serve appropriate previews in search engines and social media. Article pages in the Anthology also provide bibliographic metadata using Dublin Core, Eprints¹⁶ and Highwire Press¹⁷ schemas, optimising their integration with (academic) search engines and with automatic imports of reference managers, such as the popular open source manager Zotero¹⁸. Entry pages in the Compendium are separate entities from the dataset they describe, which is reflected in their metadata. As a result, Zotero imports produce webpage references for Compendium entries, rather than dataset or paper references.

The Compendium also renders its dataset entries in OLAC (Bird and Simons, 2001) and CMDI (Broeder et al., 2012) formats. These formats are then used to integrate the Compendium with syndicated metadata repositories, as we discuss further in section 4.2. Their schemas allow more detailed descriptions of datasets, so we include as much of each entry's information as fits with each schema. CMDI also supports a variety of profiles to describe resources of different types and modalities, as well as at different degrees of granularity. For the time being, we use a profile designed for mapping OLAC data to CMDI, but we are investigating other profiles to determine the ones most suitable for general descriptions of sign language corpora and lexica.

¹⁴<https://www.dublincore.org>

¹⁵<https://ogp.me/>

¹⁶<http://purl.org/eprint/terms>

¹⁷<https://scholar.google.no/intl/en/scholar/inclusion.html#indexing>

¹⁸<https://www.zotero.org>

4.2 Connecting resources

A core component of both our resources is to provide a wealth of external links for attribution and user guidance. Resource entries link to project pages, data sources, annotation guidelines, relevant publications, and more. Article pages link not only to conference, workshop and publisher websites, but also to their corresponding entry in the ACL Anthology.

Data entries in the sign-lang@LREC Anthology link to their more detailed counterpart in the Compendium and the Compendium in turn assists literature reviews by linking to data and project entries in the Anthology. For visitors seeking datasets that lie outside the scope of the Compendium, we also provide an overview of other catalogues of sign language data.

To help with making the datasets themselves more findable, Compendium metadata is also exposed in formats suitable for registration with syndicated metadata repositories (see also section 4.1). Since early 2023, the Compendium has been registered with the Open Language Archives Community repository (OLAC). As of May 2025, Compendium entries are also being included in the CLARIN Virtual Language Observatory (VLO). As the entries in these syndicated repositories should be understood to be descriptions of the primary resources, we take care to prioritise identifiers and links for the resources themselves and deal with references to the Compendium as a meta-information source.

4.3 Language Names and Identifiers

A component of our repositories that is notably more complex than it appears at first glance is the language index. For each language, we provide its ISO 639-3 identifier¹⁹, Glottocode (Forkel and Hammarström, 2022) and what names it is commonly known under in English and in (written) languages of its home region(s). While straightforward for many spoken languages, these matters are more complicated for sign languages.

Many sign languages have more than one name, which may originate either in their own language community or in academic literature. The history and evolution of these names is often intertwined with matters of deaf identity and the (both historic and ongoing) struggle to have sign languages recognised as independent natural languages, but also

¹⁹<https://iso639-3.sil.org/>

The Sign Language Dataset Compendium

Start | About | Corpora | Lexical Resources | Tasks | Languages | Credit

Language

Japanese Sign Language

ISO 639-3: jsl
 Glottolog: japa1238
 Acronyms: JSL, NS, NSG

English name: Japanese Sign Language
 Local names: 日本手話 (Nihon Shuwa), 日本手話言語 (Nihon Shuwa Gengo)

Corpora involving Japanese Sign Language

- Japanese Sign Language Colloquial Corpus

Lexical Resources involving Japanese Sign Language

- Asian Signbank
- SpreadTheSign

The Sign Language Dataset Compendium v1.4.0
 Contact | Imprint | Data Privacy

Figure 5: Compendium entry for Japanese Sign Language. Shows ISO 639-3 and Glottocode identifiers, followed by name information specifying acronyms, English name and Japanese name variants in Kanji and Latin transliteration, and finally the lists of corpora and lexical resources in the Compendium.

issues of ableism and academic colonialism (Batterbury et al., 2007; Bone et al., 2021; Hochgesang, 2021; Börstell, 2023). Care must therefore be taken to avoid inclusion of names that devalue their state as independent natural languages, such as names that equate them to “mimicking”, “gesturing” or mere support forms of a spoken language.

It is also common practice in both academia and signing communities to use acronyms to refer to sign languages. These should preferably be based on the community-preferred local name, though historically English-based acronyms have also been common. For example, the use of *SSL* for Swedish Sign Language has been superseded by *STS*, referencing its Swedish name *Svenskt Teckenspråk*. As an additional complication, some acronyms happen to be strongly ambiguous, especially when based on the common English “*REGION Sign Language*” pattern (e.g. *ISL* may refer to Irish-, Israeli-, Inuit-, or Indian Sign Language).

In designing the language index for our repositories we try to strike a balance between prominently displaying community-preferred names and acronyms, improving findability by listing relevant alternatives, and avoiding disrespectful names.

4.4 Synergies

Wherever possible we seek to identify ways in which efforts of one resource can be used to support another. These include adding value to users through cross-references (see section 4.2), shared literature review processes and shared information structures.

From the beginning, dataset discovery for the Compendium built on the review of sign-lang@LREC publications as a prime source of information on sign language resources (see section 3.2). We continue this practice with each new workshop, using the article review required for producing the categorial indices of the Anthology to also scan for mentions of datasets that might be suitable for the Compendium.

Where information between Anthology and Compendium overlap, we try to source them from the same structures, such as using bibliographic entries from the Anthology in the Compendium and using the same metadata for each repository’s language index (see section 4.3). Other entry types were originally built separately, due to the different needs and coverage of each repository, although work is now underway to produce flexible data structures that can serve both platforms.

Another case of synergy occurred regarding the automatic production of BibTeX entries. To ensure correct capitalisation during BibTeX conversion from title caps to sentence caps, words that should always be capitalised must be specially marked. This is a common occurrence in sign language research, as many paper titles contain language, location and resource names. While some cases of capitalisation can be detected through heuristics, other cases, language and location names in particular, are best handled by an explicit list of capitalised words. In developing such a list for the sign-lang@LREC Anthology, we used the word list of ACL Anthology as a starting point and then extended its coverage to fit the needs of our community. The expanded list was then submitted for re-integration with the ACL Anthology, resulting in improved capitalisation for 350 articles.²⁰

5 Fighting Reference Rot

A major concern in maintaining our repositories is that of *reference rot* (Klein et al., 2014). This covers the related issues of *link rot*, where a link no longer leads to the resource it once referenced, and *content drift*, where content evolved to such an extent that it no longer contains the referenced information.

For the repositories themselves, we stave off link rot by following FAIR principles. Each repository is assigned a DOI as a persistent identifier, URLs

²⁰<https://github.com/acl-org/acl-anthology/issues/953>

are kept as stable as possible and retired URLs are assigned redirects. The Compendium is also produced as a monolithic PDF document, each release of which is archived in a FAIR repository.

Dealing with reference rot of external links is a more challenging issue, and one that we have encountered regularly, especially in our work on the Compendium. Cases we have encountered included *a)* custom web domains not being renewed after the end of a project, *b)* content moving to new URIs without redirect due to website redesigns or changes to content management software, *c)* information (especially descriptions of completed projects) being moved, abbreviated or deleted entirely, *d)* dynamically generated websites failing due to broken server backends, *e)* content becoming inaccessible due to external changes, such as browsers or operating systems ceasing support for specific file formats and software types.

As our repositories are living resources, we can address some of these issues by finding new or alternative locations for the information or data in question. In other cases, the original information is lost and we must turn to web archives for help. In either case, we must first become aware that the status of a reference has changed. We also need to serve users with ways to triage issues that arise between releases. These matters we address in the following sections.




5.1 Availability and Archival Workflow

As a third party, the Compendium is not in a position to directly address the web hosting issues of other resources, but we can work towards the (partial) preservation of information. To some degree the Compendium itself represents such documentation, but to also preserve its primary sources, we must rely on the services of web archives.

One of the best known such archives is the *Wayback Machine*²¹ by the *Internet Archive*. As of time of writing, its collection reportedly contains over 928 billion web pages, including snapshots of the same page from different points in time, all of which can be viewed publicly. Archival of a web page can be triggered either by an automatic web crawl or upon user request.

Use of the Wayback Machine was a part of the editorial workflow for the Compendium from its start, helping us in recovering documentation for older resources, verifying defunct article references and

²¹<https://web.archive.org/>

Webpages	Project page: https://www.plm.uw.edu.pl/projekty/korpus-pjm/  Dataset: https://www.korpuspjm.uw.edu.pl/en 
Institution	University of Warsaw
Publications	https://www.plm.uw.edu.pl/publikacje/ 

[Go to archival copy on Internet Archive](#)

Figure 6: Excerpt of Compendium entry showing multiple external links. Each link has an archival snapshot that can be reached by clicking on the icon after the regular link. Hovering over the icon provides an explanatory tooltip.

securing pages against future loss. Having started as manual measures, performed on a per case basis, our latest release introduces an automatic workflow to consistently ascertain and ensure the archival status of external links.

Our workflow automatically iterates over the external URLs of the Compendium. For each URL, an HTTP request is sent to determine whether it is still reachable. If its availability has changed or the server rejects the request, the URL is logged for manual verification. If the archival status of the URL has not been ascertained before, an API request for archive snapshots is sent to the Wayback Machine. To avoid content drift, we select the snapshot closest to the date of inspection noted for the URL (or date of the last major revision for its entry), rather than the latest one. If no snapshot exists for the URL, creation of one is requested. Upon completion, the availability and archival information is stored with the URL as additional metadata (see the upcoming section 5.2).

This process serves to provide documentation and metadata regarding the resources described by the Compendium. The right and responsibility of providing and archiving the datasets themselves remains with their creators. Archival of pages may also fail partly or fully in individual cases. Common causes that we encountered were failure to store video materials served by third party services like Youtube and failure to store pages that dynamically serve content from a database backend.

5.2 Robust Links

Having determined the availability and archival status of our external links, there is a need to store this information and to serve it to users in an appropriate manner. For this, we build on the concept of *robust links* proposed by Klein et al. (2018)²². Robust links decorate HTML hyperlink anchors with

²²A current revision of the proposal is being worked on by Alam et al. (2025).

three new attributes that complement the existing href attribute that specifies the regular destination of a hyperlink:

1. `data-originalurl`: The original target URL, relevant when href has to be changed to fallback location.
2. `data-versiondate`: The date on which the linked content was accessed.
3. `data-versionurl`: The URLs of one or more archival snapshots.

These decorators allow us to store the archival status information obtained in section 5.1. Internally, we complement them with additional attributes to mark cases such as defunct links without backups (whose URL should nevertheless be retained for replicability) and unusable snapshots (e.g. due to broken dependencies to live databases).

These various attributes are then used during production of the repository output formats to provide links in appropriate ways. In all formats, a discrete backup link is added after the regular link (see fig. 6). In HTML outputs, anchor elements are also explicitly decorated as robust links to support processing by suitable parsers.

6 Impact

Both the Anthology and Compendium are meta-resources whose main purpose is to guide users to other resources, a task that is rarely credited explicitly. As such, their exact impact can be difficult to judge, especially for the Anthology, which until now had no associated publications that could be cited. At least one study, [Sprugnoli \(2025\)](#), explicitly names both our resources as the basis for their own survey. [Aonuki and Hall \(2024\)](#) recommend the Compendium to lecturers of linguistics classes seeking to include a diversity of sign languages in their materials.

A look at the citations of [Kopf et al. \(2021\)](#) and [Kopf et al. \(2022a\)](#) reveals additional uses for the Compendium: Most frequent is its use as a survey paper, serving as a shorthand in discussions of related work. Other publications use the Compendium's dataset inventory and its information on dataset sizes and creation periods to support observations regarding the scarcity of sign language datasets, the recent increase in number of datasets, and for size comparisons between datasets.

7 Limitations

7.1 Scope

There are certain limitations to the thematic scope of the resources described in our repositories. The Anthology is naturally limited to only catalogue content referenced by its publications. The curation criteria of the Compendium were designed to ensure a focus on resources relevant to linguistic research on language use as exhibited by signers for whom it is their language of daily life. As such it does not cover corpora that focus on script-based language production, translated or interpreted content, or on language learners and language acquisition.

This focus was also relevant for developing a consistent entry format, as the information needs in domains like machine translation or language acquisition differ noticeably from those of general sign linguistics. To assist researchers seeking materials that fall outside the scope of our collection, we provide an overview of other relevant sources of information.

Another limitation of scope is that we are unable to perform extensive quality control on the resources listed by our repositories. Both Anthology and Compendium are designed to help find potentially relevant resources, but it remains the reader's responsibility to verify that the methodological and ethical criteria of a resource make it suitable for their specific work.

7.2 Categorisation

Handcrafting categorial indices is feasible, if labour intensive, for repositories such as the `signlang@LREC` Anthology, but would be unlikely to scale to larger collections with tens of thousands of articles.

It also presents various challenges with regard to extracting required information and determining appropriate cut-offs for categorisation. Papers vary strongly in how and whether affiliations and funding are acknowledged and how well these can be mapped to a named project. Many papers also describe automatic classifiers, but an editorial decision is required to identify which might qualify as tools suitable for use by third parties. Similarly, many papers mention the use of popular editors such as ELAN, but to warrant inclusion in its entry, papers must be found to either contribute to its development or provide notable insights regarding its use.

During development of the Anthology, we also considered inclusion of a topics index, but held off on it after early tests highlighted the difficulties of consistently applying meaningful categories. Some of these matters may in future be improved by following the example of LRE Map in requesting additional information during paper submission.

7.3 Archive Availability

Our archival strategy currently relies directly on the availability of the Internet Archive Wayback Machine. Like any resource, it is exposed to a number of risks that may threaten its continued availability (Freeland, 2024). We are investigating whether additional archives may be added to our workflow to provide redundancy.

7.4 Repository Availability and Maintenance

Like the resources they describe, our repositories need to ensure their ongoing availability. Both repositories are static websites that rely only on basic and well established web standards (HTML, CSS, minimal optional JavaScript for search) without reliance on databases or content management systems. They are hosted by University of Hamburg and each assigned a resource DOI. The PDF versions of the Compendium’s releases are also archived with the university’s research data repository. To produce new releases, we use Python pipelines with a limited number of third-party dependencies. Metadata is stored using established open source text formats.

Content maintenance of the repositories is handled by us, the authors, as part of our general academic responsibilities. For the Compendium, new and changed resources are identified in the course of our involvement with the sign language resources community. This is now aided by resource creators actively seeking us out to report corrections and new releases, making content maintenance a relatively low effort. The Anthology is primarily updated every two years as part of the sign-lang@LREC workshop series event cycle and mainly builds on data already produced in the course of event organisation and proceedings publication. The only major additional effort is the maintenance of the categorial indices (see section 7.2).

Maintainer succession, while not yet an urgent issue, will also need to be addressed eventually. For the Anthology it will likely be handled as part of workshop committee recruitment. For the Compendium this remains an open question.

8 Conclusion

We have presented our work on the creation, maintenance, and ongoing development of two repositories of sign language research data. The sign-lang@LREC Anthology is a workshop series repository of sign language publications. The Sign Language Dataset Compendium is a curated metadata repository, documenting linguistic corpora, lexica and data collection tasks.

Both repositories are open and FAIR resources with rich metadata, designed to aid researchers in finding relevant works on sign languages. Different indexes group contents by language, resource type or authorship to help users focus their search. A wealth of links connects to external sources and other repositories.


We also address the risks of reference rot through a semi-automatic workflow that combines link availability checks, web archiving and robust links to harden our efforts against information loss.

The Anthology and the Compendium are living resources that are regularly updated. Should you be aware of additional relevant resources, know of information that is missing from an entry or that has changed, have spotted inaccuracies, or wish to provide us with any other feedback, please contact the Anthology team at anthology@dgs-korpus.de or the Compendium team at sldc@dgs-korpus.de.

Acknowledgements

We would like to thank Amy Isard and the anonymous reviewers for their feedback on this article. We thank Timm Lehmborg for his assistance in setting up VLO syndication.

This work has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies’ Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies’ Programme is coordinated by the Union of the Academies of Sciences and Humanities

This work was supported by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union’s Horizon 2020 research and innovation programme, grant agreement n° 101016982. 

References

- Sawood Alam, Shawn M. Jones, Martin Klein, Michael L. Nelson, and Herbert Van de Sompel. 2025. [Robustifying links](#).
- Yurika Aonuki and Kathleen Currie Hall. 2024. [Incorporating sign language phonetics & phonology exercises into the linguistics classroom](#). *The title of this volume is shorter than its contributions are allowed to be: Papers in honour of Hotze Rullmann*, pages 19–38.
- Sarah C E Batterbury, Paddy Ladd, and Mike Gulliver. 2007. [Sign language peoples as indigenous minorities: Implications for research and policy](#). *Environment and Planning A: Economy and Space*, 39(12):2899–2915.
- Steven Bird and Gary Simons. 2001. [The OLAC metadata set and controlled vocabularies](#). In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*, pages 7–18, Toulouse, France.
- Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. [Two decades of the ACL Anthology: Development, impact, and open challenges](#). In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94, Singapore. Association for Computational Linguistics.
- Tracey A. Bone, Erin Wilkinson, Danielle Ferndale, and Rodney Adams. 2021. [Indigenous and deaf people and the implications of ongoing practices of colonization: A comparison of Australia and Canada](#). *Humanity & Society*, pages 1–27.
- Carl Börstell. 2023. [Ableist language teching over sign language research](#). In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 1–10, Tórshavn, the Faroe Islands. Association for Computational Linguistics.
- Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. [CMDI: a component metadata infrastructure](#). In *Proceedings of the LREC 2012 Workshop Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources.*, pages 1–4, Istanbul, Turkey. European Language Resources Association.
- Nicoletta Calzolari, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Irene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis. 2010. [The LREC map of language resources and technologies](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 949–956, Valletta, Malta. European Language Resources Association (ELRA).
- Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, and Dimitar Shterionov. 2022. [Challenges with sign language datasets for sign language recognition and translation](#). In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 2478–2487, Marseille, France. European Language Resources Association (ELRA).
- DGS-Korpus. 2022. [730.082.677.672.551 bytes](#). *Zahl der Woche*. Union of the German Academies of Sciences and Humanities.
- Christian Federmann, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli, and Marc Schröder. 2012. [META-SHARE v2: An open network of repositories for language resources including data and tools](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3300–3303, Istanbul, Turkey. European Language Resources Association (ELRA).
- Robert Forkel and Harald Hammarström. 2022. [Glot-codes: Identifiers linking families, languages and dialects to comprehensive reference information](#). *Semantic Web*, 13(6):917–924.
- Chris Freeland. 2024. [Internet Archive and the Wayback Machine under DDoS cyber-attack](#). *Internet Archive Blogs*.
- Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. [The ACL Anthology: Current state and future directions](#). In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.
- Twan Goosen and Thomas Eckart. 2014. [Virtual Language Observatory 3.0: What's new?](#) In *Selected papers from the CLARIN 2014 Conference*, page 4, Soesterberg, Netherlands.
- Laurel L. Haak, Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. 2012. [ORCID: a system to uniquely identify researchers](#). *Learned Publishing*, 25(4):259–264.
- Thomas Hanke, Lutz König, Sven Wagner, and Silke Matthes. 2010. [DGS Corpus & Dicta-Sign: The Hamburg studio setup](#). In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 106–109, Valletta, Malta. European Language Resources Association (ELRA).
- Julie A. Hochgesang. 2021. [Open letter to Springer editors and their response](#).
- Julie A. Hochgesang, Ryan Lopic, and Emily Shaw. 2023. [W\(h\)ither the ASL corpus? considering trends in signed corpus development](#). In Ella Wehrmeyer, editor, *Advances in Sign Language Corpus Linguistics*, number 108 in Studies in Corpus Linguistics, pages 287–308. John Benjamins Publishing Company.
- Trevor Johnston. 2007. [Auslan Corpus annotation guidelines](#). Annotation convention, University of Sydney, Sydney, Australia.

- Trevor Johnston and Adam Schembri. 2006. [Issues in the creation of a digital archive of a signed language](#). In *Sustainable Data from Digital Fieldwork*, pages 7–16, Sidney, Australia. Sydney University Press.
- Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. [Scholarly context not found: One in five articles suffers from reference rot](#). *PLOS ONE*, 9(12):e115253.
- Martin Klein, Harihar Shankar, and Herbert Van de Sompel. 2018. [Robust links in scholarly communication](#). In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL '18*, pages 357–358, New York, NY, USA. Association for Computing Machinery.
- Reiner Konrad. 2012. [Sign language corpora survey](#).
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. [Overview of datasets for the sign languages of Europe](#). Project Deliverable EASIER D6.1, EASIER Consortium.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2022a. [The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association (ELRA).
- Maria Kopf, Marc Schulder, Thomas Hanke, and Sam Bigeard. 2022b. [Specification for the harmonization of sign language annotations](#). Project Deliverable EASIER D6.2, EASIER Consortium.
- Hope E. Morgan, Onno Crasborn, Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. [Facilitating the spread of new sign language technologies across Europe](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 144–147, Marseille, France. European Language Resources Association (ELRA).
- Amit Moryossef and Yoav Goldberg. 2021. [Sign language processing](#).
- Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiljevs, Gerhard Backfried, Jose Manuel Gomez-Perez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, and 17 others. 2021. [European Language Grid: A joint platform for the European language technology community](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 221–230, Online. Association for Computational Linguistics.
- Constanze H. Schmaling. 2012. [Dictionaries of African sign languages: An overview](#). *Sign Language Studies*, 12(2):236–278.
- Marc Schulder, Sam Bigeard, Thomas Hanke, and Maria Kopf. 2023. [The Sign Language Interchange Format: Harmonising sign language datasets for computational processing](#). In *Proceedings of the Eighth International Workshop on Sign Language Translation and Avatar Technology*, Rhodes, Greece. IEEE.
- Marc Schulder and Thomas Hanke. 2022. [How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages](#). In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 164–173, Marseille, France. European Language Resources Association (ELRA).
- Gary Simons and Steven Bird. 2003. [The Open Language Archives Community: An infrastructure for distributed archiving of language resources](#). *Literary and Linguistic Computing*, 18(2):117–128.
- Rachele Sprugnoli. 2025. [Current trends in online sign language dictionaries](#). *International Journal of Lexicography*, page ecaf003.
- Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. [Semantic metadata mapping in practice: the Virtual Language Observatory](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1029–1034, Istanbul, Turkey. European Language Resources Association (ELRA).
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, and 1 others. 2016. [The FAIR guiding principles for scientific data management and stewardship](#). *Scientific Data*, 3(1):9.