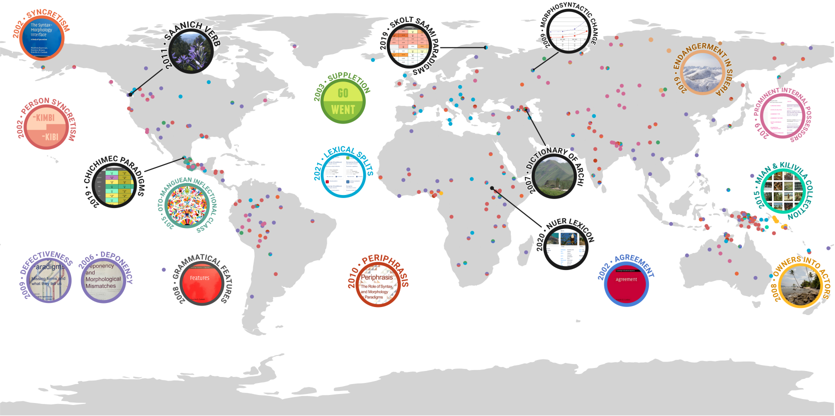
Open data: how do we get there, concretely?

Sacha Beniamine

Introduction


2002 · SYNCRETISM
The Syntax-Morphology Interface
2017 · SAANICH VERB
2019 · SKOLT SAAMI PARADIGMS
2009 · MORPHOSYNTACTIC CHANGE
2019 · ENDANGERMENT IN SIBERIA
2002 · PERSON SYNCRETISM
-KIMBI
-KIBI
2003 · SUPPLETION
GO
WENT
2019 · PROMINENT INTERNAL POSSESSORS
2019 · CHICHIMEC PARADIGMS
2021 · LEXICAL SPLITS
2007 · DICTIONARY OF ARCHI
2015 · MIAN & KILIVILA COLLECTION
2015 · OTO-MANGUEAN INFLECTIONAL CLASS
2020 · NUER LEXICON
2009 · DEFECTIVENESS
Paradigms
Missing forms and what they tell us
2006 · DEPONENCY
Deponency and Morphological Mismatches
2008 · GRAMMATICAL FEATURES
Features
2010 · PERIPHRASIS
Periphrasis
The Role of Syntax and Morphology in Paradigms
2002 · AGREEMENT
Agreement
2008 · OWNERS INTO ACTORS

- 20 years of language databases
- Looking towards the future

This talk

1. Data principles
2. What we can do when creating data
3. What we can do when publishing data
4. Conclusion

Principles and goals

Principles

Be
FAIR
Findable
Accessible
Interoperable
Reusable
and
CARE
Collective Benefit
Authority to Control
Responsibility
Ethics

Open Source

5 ★ OPEN DATA

Principles

Open Data ***1958***

Available
Editable
Re-Distributable

For everyone

Principles

Open Data ***1958***

Available
Editable
Re-Distributable

For everyone

5 ★ ***2012***

★ Available on the web
★★ Structured data
★★★ Open format
★★★★ Has URIs
★★★★★ Linked data

For everyone

Principles

Open Data ***1958***

Available
Editable
Re-Distributable

For everyone

5 ★ ***2012***

★ Available on the web
★★ Structured data
★★★ Open format
★★★★ Has URIs
★★★★★ Linked data

For everyone

FAIR ***2016***

Findable
Accessible
Inter-operable
Reusable

For machines

Principles

Open Data *1958*

Available
Editable
Re-Distributable

For everyone

5 ★ *2012*

★ Available on the web
★★ Structured data
★★★ Open format
★★★★ Has URIs
★★★★★ Linked data

For everyone

FAIR *2016*

Findable
Accessible
Inter-operable
Reusable

For machines

CARE *2012*

Collective benefit
Authority to control
Responsibility
Ethics

For Indigenous communities

Benefits for us

- Publish high quality data
- Make it as useful as possible
- Minimize the cost of maintenance

Creating data

Creating data

- Metadata FAIR
- Standards Inter-operable Reusable
- Linked data ★★★★★ Interoperable Reusable
- Validation Quality

Metadata are information

L-B 258

Ncag S T O L P E, Per
Värmlands natur och värmländskt arbetsliv.
Föredrag.

1934, 10.1 (Karlstad).

Stolpe, Per.

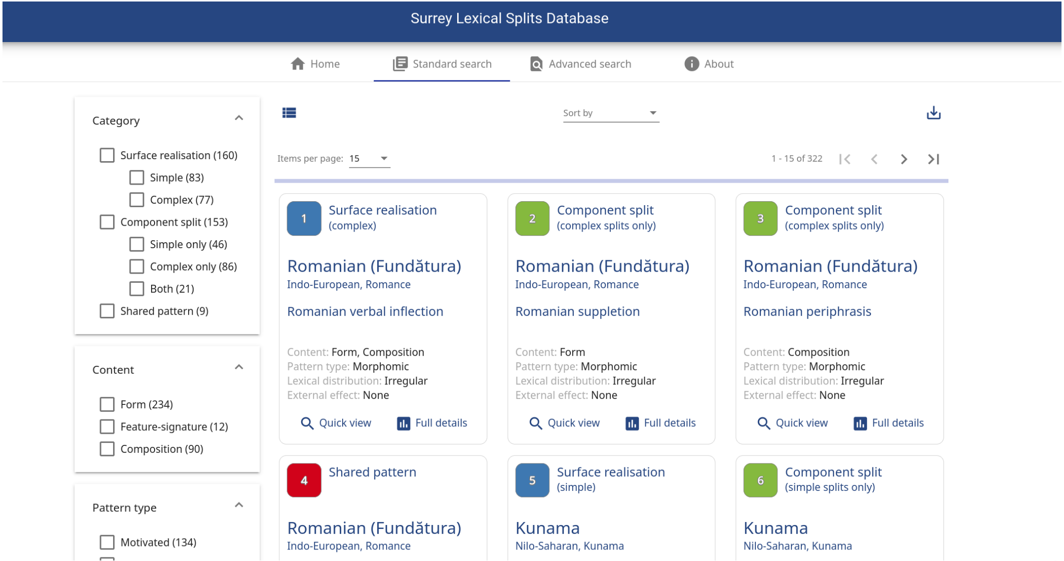
6'45

- About a dataset
  - Authors, Contributors
  - title
  - Citation
  - Grants which funded it
  - How is it related to other datasets
  - Sources
  - Date of publication

Metadata are information

- About a dataset
- About its structure
  - What conventions were used?
  - How were missing data written?
  - What is each table for?
  - What is the content in each column? Should the content be numbers, text, iso codes, references to documents, etc.

Metadata: example



Surrey Lexical Splits Database

Home | Standard search | Advanced search | About

Category

□ Surface realisation (160)
- □ Simple (83)
- □ Complex (77)

□ Component split (153)
- □ Simple only (46)
- □ Complex only (86)
- □ Both (21)

□ Shared pattern (9)

Content

□ Form (234)
□ Feature-signature (12)
□ Composition (90)

Pattern type

□ Motivated (134)

Sort by

Items per page: 15

1 - 15 of 322

1 Surface realisation (complex)

Romanian (Fundătura)
Indo-European, Romance

Romanian verbal inflection

Content: Form, Composition
Pattern type: Morphomic
Lexical distribution: Irregular
External effect: None

Quick view | Full details

2 Component split (complex splits only)

Romanian (Fundătura)
Indo-European, Romance

Romanian suppletion

Content: Form
Pattern type: Morphomic
Lexical distribution: Irregular
External effect: None

Quick view | Full details

3 Component split (complex splits only)

Romanian (Fundătura)
Indo-European, Romance

Romanian periphrasis

Content: Composition
Pattern type: Morphomic
Lexical distribution: Irregular
External effect: None

Quick view | Full details

4 Shared pattern

Romanian (Fundătura)
Indo-European, Romance

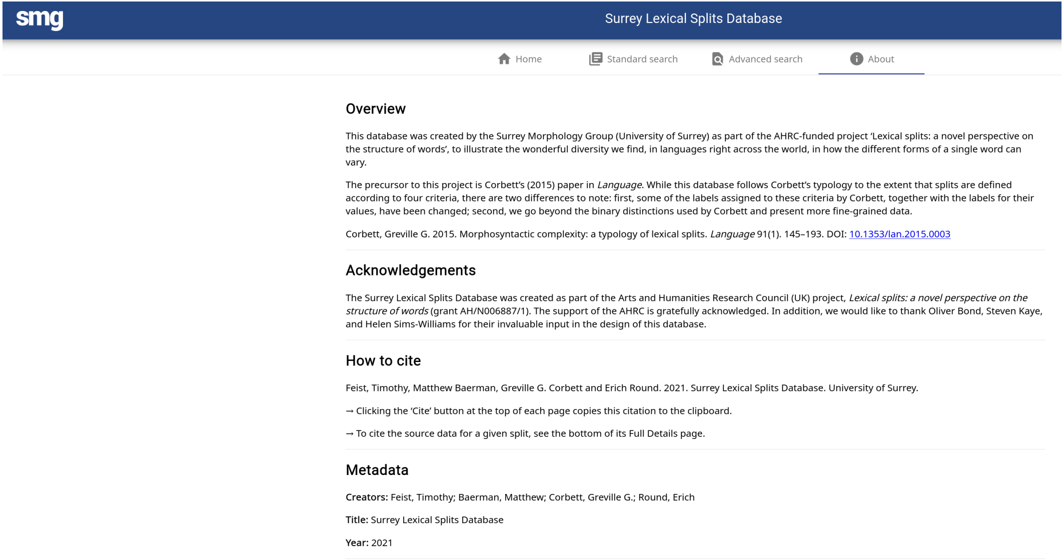
5 Surface realisation (simple)

Kunama
Nilo-Saharan, Kunama

6 Component split (simple splits only)

Kunama
Nilo-Saharan, Kunama

Metadata: example



Overview

This database was created by the Surrey Morphology Group (University of Surrey) as part of the AHRC-funded project ‘Lexical splits: a novel perspective on the structure of words’, to illustrate the wonderful diversity we find, in languages right across the world, in how the different forms of a single word can vary.

The precursor to this project is Corbett’s (2015) paper in *Language*. While this database follows Corbett’s typology to the extent that splits are defined according to four criteria, there are two differences to note: first, some of the labels assigned to these criteria by Corbett, together with the labels for their values, have been changed; second, we go beyond the binary distinctions used by Corbett and present more fine-grained data.

Corbett, Greville G. 2015. Morphosyntactic complexity: a typology of lexical splits. *Language* 91(1). 145–193. DOI: 10.1353/lan.2015.0003

Acknowledgements

The Surrey Lexical Splits Database was created as part of the Arts and Humanities Research Council (UK) project, *Lexical splits: a novel perspective on the structure of words* (grant AH/N006887/1). The support of the AHRC is gratefully acknowledged. In addition, we would like to thank Oliver Bond, Steven Kaye, and Helen Sims-Williams for their invaluable input in the design of this database.

How to cite

Feist, Timothy, Matthew Baerman, Greville G. Corbett and Erich Round. 2021. Surrey Lexical Splits Database. University of Surrey.

→ Clicking the ‘Cite’ button at the top of each page copies this citation to the clipboard.

→ To cite the source data for a given split, see the bottom of its Full Details page.

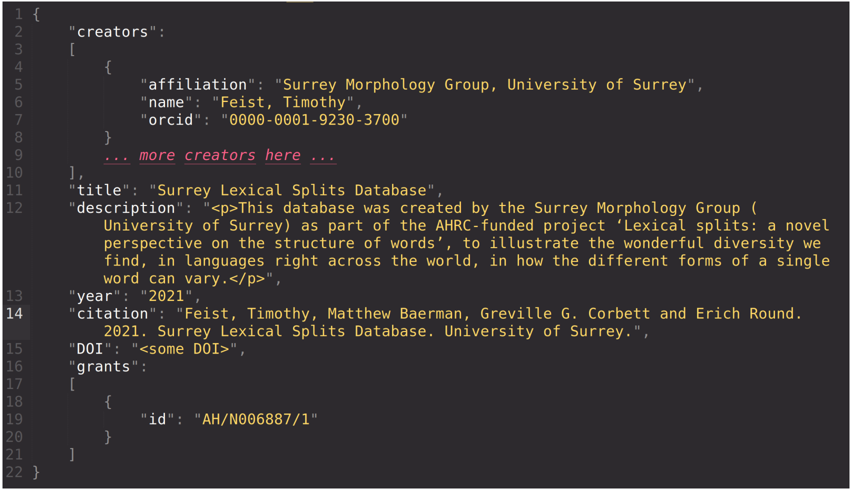
Metadata

Creators: Feist, Timothy; Baerman, Matthew; Corbett, Greville G.; Round, Erich

Title: Surrey Lexical Splits Database

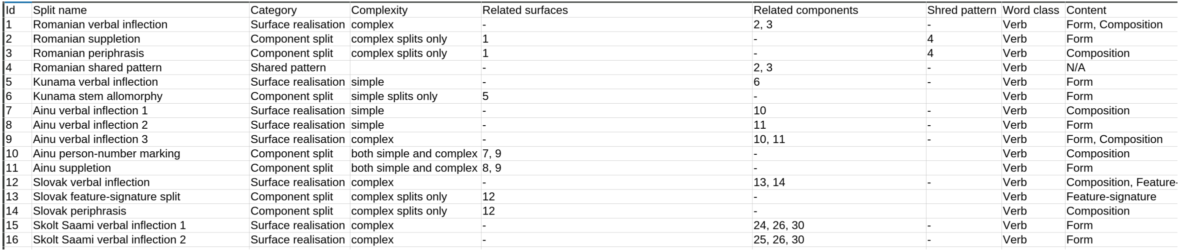
Year: 2021

Metadata: example



```
{
    "creators":
    [
        {
            "affiliation": "Surrey Morphology Group, University of Surrey",
            "name": "Feist, Timothy",
            "orcid": "0000-0001-9230-3700"
        }
        ... more creators here ...
    ],
    "title": "Surrey Lexical Splits Database",
    "description": "<p>This database was created by the Surrey Morphology Group (
        University of Surrey) as part of the AHRC-funded project ‘Lexical splits: a novel
        perspective on the structure of words’, to illustrate the wonderful diversity we
        find, in languages right across the world, in how the different forms of a single
        word can vary.</p>",
    "year": "2021",
    "citation": "Feist, Timothy, Matthew Baerman, Greville G. Corbett and Erich Round.
        2021. Surrey Lexical Splits Database. University of Surrey.",
    "DOI": "<some DOI>",
    "grants":
    [
        {
            "id": "AH/N006887/1"
        }
    ]
}
```

Metadata: example

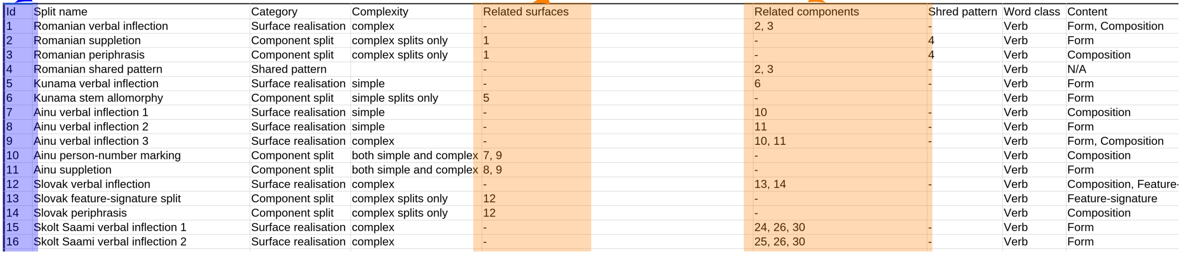


Id	Split name	Category	Complexity	Related surfaces	Related components	Shred pattern	Word class	Content
1	Romanian verbal inflection	Surface realisation	complex	-	2, 3	-	Verb	Form, Composition
2	Romanian suppletion	Component split	complex splits only	1	-	4	Verb	Form
3	Romanian periphrasis	Component split	complex splits only	1	-	4	Verb	Composition
4	Romanian shared pattern	Shared pattern		-	2, 3	-	Verb	N/A
5	Kunama verbal inflection	Surface realisation	simple	-	6	-	Verb	Form
6	Kunama stem allomorphy	Component split	simple splits only	5	-		Verb	Form
7	Ainu verbal inflection 1	Surface realisation	simple	-	10	-	Verb	Composition
8	Ainu verbal inflection 2	Surface realisation	simple	-	11	-	Verb	Form
9	Ainu verbal inflection 3	Surface realisation	complex	-	10, 11	-	Verb	Form, Composition
10	Ainu person-number marking	Component split	both simple and complex	7, 9	-		Verb	Composition
11	Ainu suppletion	Component split	both simple and complex	8, 9	-		Verb	Form
12	Slovak verbal inflection	Surface realisation	complex	-	13, 14	-	Verb	Composition, Feature-
13	Slovak feature-signature split	Component split	complex splits only	12	-		Verb	Feature-signature
14	Slovak periphrasis	Component split	complex splits only	12	-		Verb	Composition
15	Skolt Saami verbal inflection 1	Surface realisation	complex	-	24, 26, 30	-	Verb	Form
16	Skolt Saami verbal inflection 2	Surface realisation	complex	-	25, 26, 30	-	Verb	Form

Metadata: example

row IDs

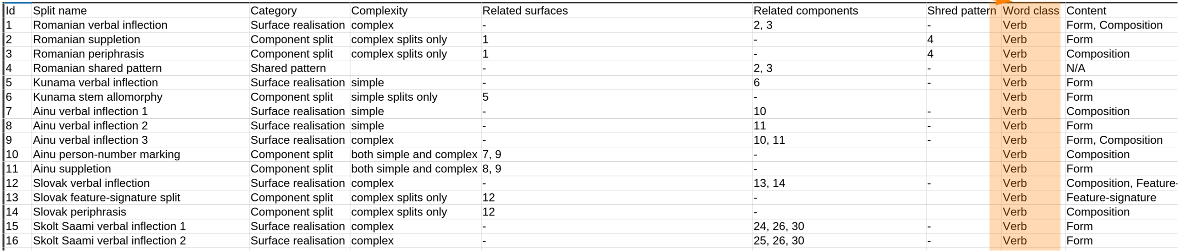
References to row IDs



Id	Split name	Category	Complexity	Related surfaces	Related components	Shred pattern	Word class	Content
1	Romanian verbal inflection	Surface realisation	complex	-	2, 3	-	Verb	Form, Composition
2	Romanian suppletion	Component split	complex splits only	1	-	4	Verb	Form
3	Romanian periphrasis	Component split	complex splits only	1	-	4	Verb	Composition
4	Romanian shared pattern	Shared pattern		-	2, 3	-	Verb	N/A
5	Kunama verbal inflection	Surface realisation	simple	-	6	-	Verb	Form
6	Kunama stem allomorphy	Component split	simple splits only	5	-		Verb	Form
7	Ainu verbal inflection 1	Surface realisation	simple	-	10	-	Verb	Composition
8	Ainu verbal inflection 2	Surface realisation	simple	-	11	-	Verb	Form
9	Ainu verbal inflection 3	Surface realisation	complex	-	10, 11	-	Verb	Form, Composition
10	Ainu person-number marking	Component split	both simple and complex	7, 9	-		Verb	Composition
11	Ainu suppletion	Component split	both simple and complex	8, 9	-		Verb	Form
12	Slovak verbal inflection	Surface realisation	complex	-	13, 14	-	Verb	Composition, Feature-
13	Slovak feature-signature split	Component split	complex splits only	12	-		Verb	Feature-signature
14	Slovak periphrasis	Component split	complex splits only	12	-		Verb	Composition
15	Skolt Saami verbal inflection 1	Surface realisation	complex	-	24, 26, 30	-	Verb	Form
16	Skolt Saami verbal inflection 2	Surface realisation	complex	-	25, 26, 30	-	Verb	Form

Metadata: example

One of ”Verb” or ”Noun”

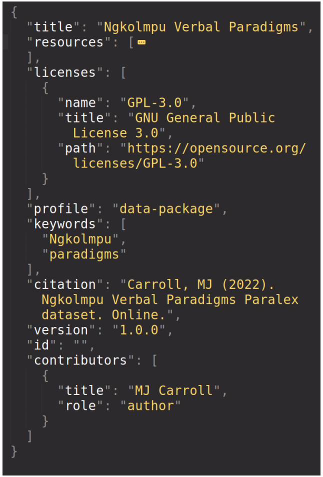


Id	Split name	Category	Complexity	Related surfaces	Related components	Shred pattern	Word class	Content
1	Romanian verbal inflection	Surface realisation	complex	-	2, 3	-	Verb	Form, Composition
2	Romanian suppletion	Component split	complex splits only	1	-	4	Verb	Form
3	Romanian periphrasis	Component split	complex splits only	1	-	4	Verb	Composition
4	Romanian shared pattern	Shared pattern		-	2, 3	-	Verb	N/A
5	Kunama verbal inflection	Surface realisation	simple	-	6	-	Verb	Form
6	Kunama stem allomorphy	Component split	simple splits only	5	-		Verb	Form
7	Ainu verbal inflection 1	Surface realisation	simple	-	10	-	Verb	Composition
8	Ainu verbal inflection 2	Surface realisation	simple	-	11	-	Verb	Form
9	Ainu verbal inflection 3	Surface realisation	complex	-	10, 11	-	Verb	Form, Composition
10	Ainu person-number marking	Component split	both simple and complex	7, 9	-		Verb	Composition
11	Ainu suppletion	Component split	both simple and complex	8, 9	-		Verb	Form
12	Slovak verbal inflection	Surface realisation	complex	-	13, 14	-	Verb	Composition, Feature-
13	Slovak feature-signature split	Component split	complex splits only	12	-		Verb	Feature-signature
14	Slovak periphrasis	Component split	complex splits only	12	-		Verb	Composition
15	Skolt Saami verbal inflection 1	Surface realisation	complex	-	24, 26, 30	-	Verb	Form
16	Skolt Saami verbal inflection 2	Surface realisation	complex	-	25, 26, 30	-	Verb	Form

What are metadata for

- Giving context to machines (and humans !)
- Ensure unambiguous interpretation
- Finding data
- Filtering data
- Manipulating & transforming data
- Validating data

How to add metadata



```
{
  "title": "Ngkolmpu Verbal Paradigms",
  "resources": [⋯
  ],
  "licenses": [
    {
      "name": "GPL-3.0",
      "title": "GNU General Public
        License 3.0",
      "path": "https://opensource.org/
        licenses/GPL-3.0"
    }
  ],
  "profile": "data-package",
  "keywords": [
    "Ngkolmpu",
    "paradigms"
  ],
  "citation": "Carroll, MJ (2022).
    Ngkolmpu Verbal Paradigms Paralex
    dataset. Online.",
  "version": "1.0.0",
  "id": "",
  "contributors": [
    {
      "title": "MJ Carroll",
      "role": "author"
    }
  ]
}
```

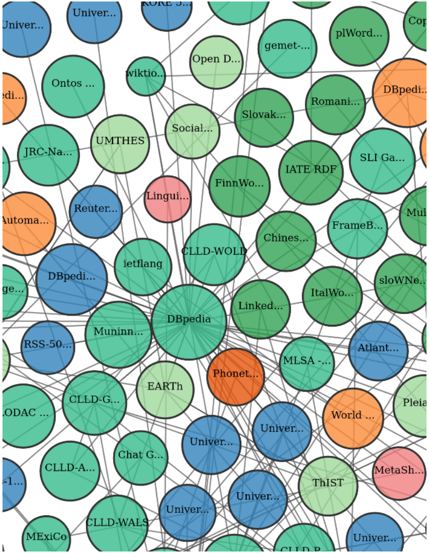
- Usually as a separate file
- machine readable format: `json`, `xml`, `yaml`
- Usually generated (forms or scripts)

Standards

- For data points
    - The metric system
    - Leipzig glossing rules
    - ISO 639 languages
    - ISO 3166 countries
- For datasets
    - Text Encoding Initiative (TEI)
    - CLDF
    - Various standards for corpora

- For the metadata
    - Frictionless
    - Dublin Core
    - CMDI

What is linked data



- Give resolvable URIs (identifiers) to data points
- Use URIs to link your data to other data
- Now the data is part of a network

◀ https://linguistic-lod.org/

What is linked data for

- Standardizing terms by linking to catalogs
- Inter-operability
- Re-usability

How to create linked data

- Provide URIs for your data points
- Find relevant vocabularies for expressing terms
- Find related entries in other databases
- Use URI Links instead of just terms
- Declare it in the metadata

What is validation

- Problems of manual inputs
- Validating the structure (syntax)
- Validating linked data
- Validating content and types
- Testing
- How?
  - Use online validators
  - Hire an engineer

Publishing data

Publishing data

- Documentation Reusable
- DOIs Findable Accessible ★★★★
- License Open Reusable
- Download in structured, open formats ★★★ Inter-operable Reusable
- Archiving Findable

Licenses

Define how data can be accessed, shared, distributed

GPL-v3

Share-alike
State changes
Attribution

For software

CC BY-SA 4.0

SA: Share-alike
BY: Attribution

For data

License pickers:

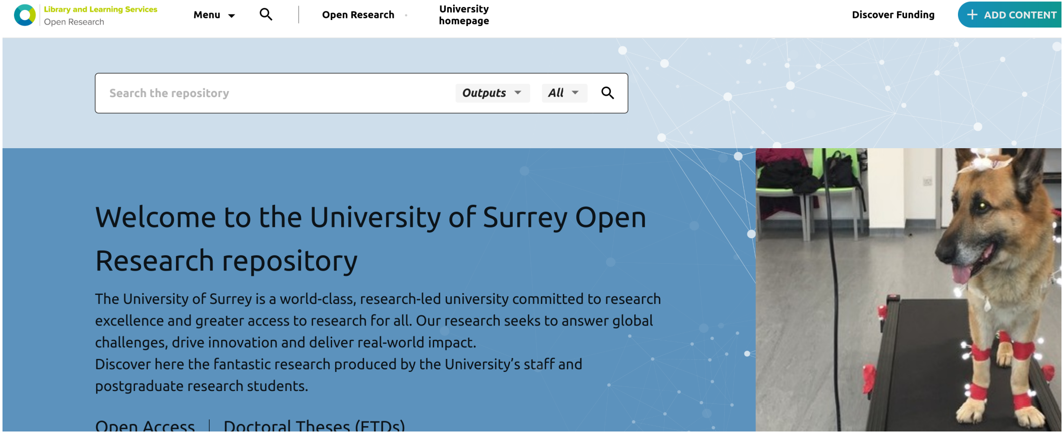
- `https://choosealicense.com/`
- `https://creativecommons.org/choose/`

Downloads

- Why?
  - Websites are show-cases, but not archives
  - Quantitative work requires downloads
  - To aggregate, modify, etc

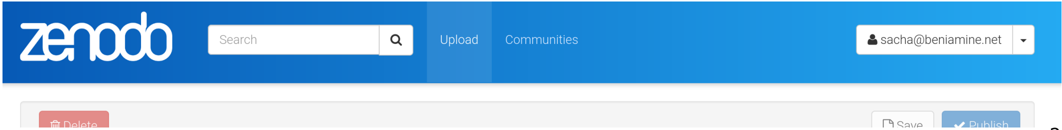
- How
  - Full data (not just a query)
  - With license & metadata
  - In open formats

Archives



Welcome to the University of Surrey Open Research repository

The University of Surrey is a world-class, research-led university committed to research excellence and greater access to research for all. Our research seeks to answer global challenges, drive innovation and deliver real-world impact.
Discover here the fantastic research produced by the University's staff and postgraduate research students.



Data check list

1. **Linked data**
   - □ Has a DOI
   - □ Defines URIs (if relevant)
   - □ Uses linked identifiers
2. **Standards**
   - □ For data points
   - □ For the download files
   - □ For the metadata
3. **Validation**
   - □ For the data format
   - □ For the content
4. **Metadata**
   - □ about the dataset
   - □ about its format & content
   - □ license
   - □ plain text doc
5. **Downloads**
   - □ entirely, with metadata
   - □ In a structured format
   - □ In an open format
   - □ From an archival site

Conclusion

- Good data practices intersect
- Following these benefits everyone
- Going further:
  - Updating data (versioning systems, continuous validation)
  - Tracking citations
  - Summaries across DB