Print
Wednesday, 30 August 2023, 3:52 PM

Site: **UPSKILLS**
Course: **Introduction to Language Data: Standards and Repositories (Introduction to Language Data: Standards and Repositories)**
Glossary: **Key concepts**

A

annotation

Definition

Description or analysis that is associated with "raw data" or with other annotations. Examples include transcription, glossing, and part-of-speech annotation.

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

C

CMDI

Component Metadata Infrastructure (CMDI)

A framework that researchers can use to combine several metadata components into a self-defined schema that suits their particular needs.

Note:

It can be used to give a detailed metadata description of a resource and can be customized by using only the metadata elements relevant to the resource. CMDI is compatible with other existing frameworks such as IMDI, OLAC, Dublin Core or the TEI header.

Example:
The Virtual Language Observatory (VLO) offers a metadata-based search on over a million language resources described with CMDI metadata.

Source:
CLARIN - https://www.clarin.eu/content/component-metadata

curation

Definition

The action of maintaining, preserving, and adding value to digital research data throughout its lifecycle.

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

D

data archive

Definition

a deposition database (also called a repository) collecting primary data from data producers (scientists, machines, etc.) arisen from different sources (experiments, observations, simulations, etc.)

Source

SSHOC Multilingual Data Stewardship Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshocterm/data_archive_20

Tags: archive repository

data archiving

Definition

The action of transferring data to a resource provider, e.g. a. repository or a data centre, all while complying with any documented guidance, policies, or legal requirements.

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

data citation

Definition

The practice of providing an identifying reference to data in a similar way that researchers routinely include a bibliographic reference to published resources.

Source
SSHOC Multilingual Data Stewardship Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshoctermdata_citation_30

Learn more

Watch this YouTube video to learn more about persistent identifiers and data citation:

.

data curation

Definition

The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for discovery and reuse.

Source

SSHOC Multilingual Data Stewardship Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshocterm/data_curation_37

Tags: data curation | data discovery | data reuse

data deposit

Definition

The process by which data is stored in a data archive.

Source

SSHOC Multilingual Data Stewardship Terminology

https://vocabs.sshopencloud.eu/vocabularies/sshocterm/data_deposit_39

Tags: data archive | data deposit | data deposition | deposit data

data journal

Definition

A (peer-reviewed) journal designed to comprehensively document and publish deposited datasets and to facilitate their online exploration.

Source

The Research Data Journal for the Humanities and Social Sciences

Tags: datasets | data | sharing | publishing

data management

Definition

A general name for the many tasks involved with proper care for research data (e.g. data collection, storage, organization, analysis, sharing, and preservation.)

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

Tags: research data management

data publication

Definition

The release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way.

Source
SSHOC Multilingual Data Stewardship Terminology
https://vocabs.sshopencloud.eu/vocabularies/sshocterm/data_publication_93

Tags: data publication | data publishing

data sharing

Definition

The practice of making data available for reuse. This may be done, for example, by depositing the data in a repository, through data publication.

Source
SSHOC Multilingual Data Stewardship Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshocterm/data_sharing_114

Tags: sharing | repository

dataset

Definition

A collection of data, published or curated by a single agent, and available for access or download in one or more formats.

Source

SSHOC Multilingual Data Stewardship Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshoctermdata_set_113

Learn more

A digital dataset might comprise a single time such as a spreadsheet of numerical data, or it might be larger, comprising a collection of related items such as spreadsheets, images, and many other types of data.

Digital Object Identifier (DOI)

Definition

A persistent identifier used to identify objects uniquely, standardized by the International Organization for Standardization (ISO). The developer and administrator of the DOI system is the International DOI Foundation (IDF), which introduced it in 2000.

Source
SSHOC Multilingual Data Stewardship Terminology,
https://vocabs.sshopencloud.eu/vocabularies/sshoctermdigital_object_identifier_140

Learn more

DOIs are also handles, but with additional policies involved (such as specific DataCite metadata) and a specific landing page for the dataset. DOI is a proper service that is used in particular by publishing companies, but its independent business model is not acceptable by many research organisations. DOIs are registered by DataCite.

Dublin Core

Definition

A widely used international metadata standard used to describe a resource in repositories at a very basic level: e.g. author, description and language.

Acronym: DC

Learn more

It is managed by the Dublin Core Metadata Initiative (DCMI) and it consists of a total of 18 elements, which can be used to describe a resource in terms of its content, version and intellectual property: *Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, Audience, Provenance and RightsHolder.*

E

European Open Science Cloud

Definition

A federated and open multi-disciplinary environment where EU researchers can publish, find and reuse data, tools and services for research, innovation and educational purposes.

Source

European Open Science Cloud (EOSC) | European Commission (europa.eu)

Learn more

For a general introduction to EOSC, follow this training by Library Carpentry: The European Open Science Cloud (EOSC) (librarycarpentry.org)

F

FAIR Principles

Definition

A collection of guidelines that can be applied to improve the **Findability**, **Accessibility**, **Interoperability**, and **Reusability** of data objects.

Note

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data*. They are described in detail on the website of GO-FAIR organisation.

Source

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Learn more

1. Watch this knowledge clip to learn more about the FAIR data principles and how to make your research data FAIR by depositing it in a research data repository:

2. Watch this video by Dieter Van Uytvanck to learn how the CLARIN research infrastructure makes the deposited language resources FAIR.

federated login

Definition

An authentication method that allows users to access the infrastructure or services with their institutional credentials.

Source

Federated identity | CLARIN ERIC

federated research infrastructure

Definition

All resources and their metadata are equally accessible and searchable throughout the infrastructure (a single sign-on), irrespective of their physical location.

Learn more

Federated identity | CLARIN ERIC

H

Handle

A persistent identifier that consists of a prefix and a suffix and is assigned to digital objects.

Note

The prefix is a numerical code indicating the institution or organisation that assigned the handle. The Handle system is developed by the Corporation for National Research Initiatives.

L

language resource

Definition

Speech and language data type in machine-readable form, as well as tools and services for the processing of language data.

Notes

Examples of language resources are: written or spoken corpora and lexica, multi-modal resources, grammars, terminology or domain-specific databases and dictionaries, ontologies, multimedia databases, etc.

Following a longstanding tradition (Godfrey & Zampolli 1997), language resources may also include software tools for the preparation, collection, management, or use of other resources, e.g. corpus management and exploration systems, OCR systems, NLP pipelines, speech processing systems, machine translation systems, environments for manual annotation and evaluation.

Learn more

Explore the language resources made available in the CLARIN research infrastructure.

M

metadata

Definition

A unique and stable denomination (reference) of a digital resource (e.g. research data) through the allocation of a code that can be persistently and explicitly referenced on the internet.

Source

SSHOC Multilingual Data Stewardship Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshocterm/en/page/metadata_158

Learn more

Literally, metadata means "data about data", i.e. data that defines and describes the characteristics of other data, used to improve both business and technical understanding of data and data-related processes. Common types:

- *Business metadata* includes the names and business definitions of subject areas, entities and attributes, attribute data types and other attribute properties, range descriptions, valid domain values and their definitions.
- *Technical metadata* includes physical database table and column names, column properties, and the properties of other database objects, including how data is stored.
- *Process metadata* is data that defines and describes the characteristics of other system elements (processes, business rules, programs, jobs, tools, etc.).
- *Data stewardship metadata* is data about data stewards, stewardship processes and responsibility assignments.

Source for the note: Scholars Portal

Learn more

Watch the video below explaining the difference between data and metadata:

.

metadata schema

Definition

Logical plan showing the relationships between metadata elements.

Source

SSHOC Multilingual Data Stewardship
Terminology, https://vocabs.sshopencloud.eu/vocabularies/sshocterm/en/page/metadata_schema_162

O

Open Language Archives Community (OLAC)

Definition

A metadata standard used to describe resources in Social Sciences.

Notes

It extends the Dublin Core fields with additional fields to describe linguistic data types that come from different archives.

Source

The Open Language Archives Community (OLAC): http://www.language-archives.org/?msclkid=0930df24a63511ec941c8e7184f5ab1d

Open Lexicon Interchange Format (OLIF)

Definition

An open, XML-compliant standard for the exchange of terminological and lexical data.

Note

Although originally intended as a means for the exchange of lexical data between proprietary machine translation lexicons, it has evolved into a more general standard for terminology exchange.

Source

http://www.olif.net/

Open Researcher and Contributor ID (ORCID)

Definition

A type of persistent identifier that researchers can obtain for free to distinguish themselves from other researchers and connect their personal information to their affiliations, grants, publications, and peer review.

Note

You can use an ORCID ID to log in to research data repositories, e.g. Zenodo.

Source

ORCID

Learn more

Watch the following videos to learn more about ORCIDs and how you can get one.

- What is an ORCID? - Source ORCID on Vimeo
- A quick tour of the ORCID record? - - Source ORCID on Vimeo

Open Science

Definition

The movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of society, amateur or professional.

Notes

Open science is transparent and accessible knowledge that is shared and developed through collaborative networks. It encompasses practices such as publishing open research, campaigning for open access, encouraging scientists to practice open-notebook science, broader dissemination and engagement in science and generally making it easier to publish, access, and communicate scientific knowledge.

Source

Open science - Wikipedia

Learn more

Watch this video to learn more about open science.

P

Persistent Identifier (PID)

Definition

An electronic identification referring to or citing electronic documents, files, resources, resource collections such as books, articles, papers, images etc.

Note

In contrast to URL, which can be changed if the resources move to other servers or to other organisations, PID does not change and always remains the same. That allows the necessary resources to be reliably identifiable and to remain accessible at all times.

Examples of PIDs: Handle (HDL), Digital Object Identifier (DOI), Uniform Resource Name (URN), the persistent URL (PURL), National Bibliography Numbers (NBNs).

The CLARIN repositories use the Handle system to assign, manage and resolve persistent identifiers for digital objects.

Source

CLARIN Standards Information System - PISA standard

Learn more

What are Persistent Identifiers and why use them?

R

reproducible research

Definition

Research is defined as reproducible when the published results can be replicated using the documented data, code, and methods employed by the author or provider without the need for any additional information or need to communicate with the author or provider.

Note

Reproducible research is generally seen as a marker of good research design.

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

research data

Definition

The various bits of evidence that arise during research and upon which conclusions, analyses, observations, generalizations, etc., can be made.

Note

Data in language-related activities means all samples of language, e.g. recordings and written language, words, sentences, verbal art, storytelling, oration etc.

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

Tags: research data management

research data lifecycle

Definition

A model that illustrates the stages of data management and describes how data flow through a research project from start to finish.

Source

SHHOC Multilingual Data Stewardship Terminology

https://vocabs.sshopencloud.eu/vocabularies/sshocterm/en/page/research_data_lifecycle_187

Learn more

Watch this video Research data management — UK Data Service to learn more about the 6 stages in research data management.

To understand the current practices in the linguistic data lifecycle, read this chapter by Eleanor Mattern, 2022. "The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management", The Open Handbook of Linguistic Data Management, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B.

Tags: research data | research data management

Research Data Management (RDM)

Definition

The practice of creating, maintaining, and preserving digitally created data.

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

research data repository

Definition

A database or virtual archive that is established to collect, disseminate and preserve scientific output.

Note

Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats, which are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis. Common types of repositories are general, discipline-specific and institutional,

Source

Gabber, Shirley, Danielle Yarbrough, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2022. Linguistic Data Management: Online companion course to The Open Handbook of Linguistic Data Management. Website: https://sites.google.com/hawaii.edu/linguisticdatamanagement/ [21/03/2022].

Learn more

1. Watch the video below to understand how certified digital repositories contribute to making and keeping research data findable, accessible, interoperable and reusable (FAIR).

S

Segmentation Rules eXchange (SRX)

Definition

A standard that enhances the TMX standard so that translation memory data that is exchanged between applications can be used more effectively

Note

Developed by OASIS.

single-sign on access

an authentication method that allows users to sign in using one set of credentials to multiple independent software applications

T

TEI

Text Encoding Initiative (TEI)

A widely used standard in the Humanities for the markup of electronic texts, ranging from corpora like the BNC to poetry.

Note

The encoding relies on SGML or XML. Metadata can be embedded in the header of a TEI file, which generally contains fields that correspond to a bibliographic record (e.g. title, distributor). TEI header elements are widely known also in the Language Resource Technology domain and are used in a number of projects to characterize resources.

Source
Text Encoding Initiative (TEI): https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

TermBase eXchange (TBX)

Definition

A standard that allows for the interchange of terminology data including detailed lexical information.

Note

Developed by LISA and revised and republished by ISO 30042.The framework for TBX is provided by three ISO standards: ISO 12620, ISO 12200 and ISO 16642.

Translation Memory eXchange (TMX)

Definition

An XML-based standard for importing and exporting translation memories created by computer-aided translation and localization tools.

Source

CLARIN Standards Information System, https://clarin.ids-mannheim.de/standards/views/view-spec.xq?id=SpecTMX

Learn more

The TMX specification is defined in two parts:

- "A specification of the format of the container (the higher-level elements that provide information about the file as a whole and about entries). In TMX, an entry consisting of aligned segments of text in two or more languages is called a Translation Unit (the <tu> element).
- A specification of a low-level meta-markup format for the content of a segment of translation-memory text. In TMX, an individual of translation-memory text in a particular language is denoted by a seg element." See the TMX 1.4b Specification for more details.

TRUST Principles

Definition

A set of guiding principles that allow data repositories to “demonstrate that they are reliable and capable of appropriately managing the data they hold (Lin et al., 2020).

Note

These principles are **T**ransparency, **R**esponsibility, **U**ser focus, **S**ustainability and **T**echnology.

Source

Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

U

Universal Terminology eXchange (UTX)

Definition

A simple, standardized glossary format that can be easily shared and reused across various tools.

Source

AAMT, https://www.aamt.info/english/utx/

X

XML Localisation Interchange File Format (XLIFF)

Definition

An XML-based bitext format created to standardize the way localizable data are passed between and among tools during a localization process and a standard format for CAT tool exchange.

Note

XLIFF is the preferred way of exchanging data in XML format in the translation industry.

Source

Wikipedia, https://en.wikipedia.org/wiki/XLIFF