Data Citation:
A guide to best practice

October 2021

Publications Office
of the European Union

Data Citation:
A guide to best practice

October 2021

This document was prepared for the European Commission, however it only reflects the views of the authors. Neither the European Commission nor any person acting on its behalf is liable for any consequence stemming from the reuse of this publication or the information contained therein, or for the content of the external sources, including external websites, referenced in this publication.

For more information:

OP.C.4
Publications Office of the European Union

2, rue Mercier
L-2985 Luxembourg
LUXEMBOURG

OP-DATA-EUROPA-EU@publications.europa.eu

Author: Paul Jessop

Printed by the Publications Office of the European Union in Luxembourg

The European Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2022

This publication is intended for information purposes only. It must be accessible free of charge.

This document is the deliverable for the ‘Data citation guidelines’ project (funding PA LOD).

Print	ISBN 978-92-78-42672-9	doi:10.2830/259496	OA-03-21-397-EN-C
PDF	ISBN 978-92-78-42673-6	doi:10.2830/59387	OA-03-21-397-EN-N

Contents

Foreword 7

Data in the European Union 7

What is citation? 8

What is in this guide? 9

Who is this guide for? 9

Acknowledgements 10

1. PART ONE – Citing data is important 11

1.1. Why cite? 11

1.1.1. Credit 11

1.1.2. Transparency 12

1.1.3. Integrity 12

1.1.4. Reproducibility 12

1.1.5. Reuse 13

1.1.6. Text mining 14

1.2. Specifying the cited data 14

1.2.1. Precision in the source 14

1.2.2. Specification of the target 14

1.2.3. Precision in the target 15

1.2.4. Characteristics of persistent identifiers 15

1.2.5. Types of persistent identifier 16

1.2.6. The Handle System 16

1.2.7. The Digital Object Identifier system 17

1.2.8. The Publications Office of the European Union and persistent identifiers 17

1.2.9. Custom identifiers 18

1.2.10. Using available persistent identifiers 19

1.2.12. Local identification schemes 19

1.2.13. Link rot and broken links 20

1.2.14. Other persistent identifiers in citations 20

1.3. Data that changes 21

1.3.1. Continuing datasets 21

1.3.2. Dynamic datasets 22

1.3.2.1. Snapshots 23

1.3.2.2. Citing queries 23

1.3.2.3. Versioning and citation 24

1.4. Creating good citations 25

1.4.1. Authoring tools 25

1.4.2. Citation management tools 25

1.5. Rights and licences 26

1.5.1. Licences and citation 26

1.5.2. Data published by EU institutions, bodies and agencies 26

1.5.3. Unpublished data 27

1.5.4. Informing the user about reuse access terms 28

1.5.5. Meeting the CC BY requirement 28

1.5.6. Privacy and the general data protection regulation 28

1.6. Data mining and data harvesting 29

1.7. Publishing data for good citation 29

1.7.1. Availability 30

1.7.2. Local hosting 30

1.7.3. Data repository 30

1.7.4. Precision 31

1.7.5. Getting a PID for a dataset 31

1.7.6. Persistence 32

1.7.7. 'Cite this dataset as' 32

2. PART TWO – How to cite data 33

2.1. The components of a data citation 33

2.1.1. Authors 33

2.1.2. Corporate author 34

2.1.3. Personal author(s) 35

2.1.4. Dataset title 35

2.1.5. Version/edition 36

2.1.6. Publisher 36

2.1.7. Publication date 37

2.1.8. Date of citation 37

2.1.9. Persistent identifier 38

2.2. The citation itself 38

2.2.1. Editorial styles........ 38

2.2.2. Citation formats and examples 40

2.2.2.1. Journalistic 40

2.2.2.2. Inline 40

2.2.2.3. Caption 41

2.2.2.4. Footnote or reference-list entry 42

2.2.2.5. Citation anchor – Interinstitutional Style Guide numbered note style 43

2.2.2.6. Citation anchor – author–date 44

2.3. Dates in citations 45

2.3.1. *Interinstitutional Style Guide* date and time format 45

2.3.2. Date and format compliance with DCAT-AP (xsd:date or xsd:dateTime) 46

2.3.3. Time zones 46

2.3.4. International standard date and time examples 47

2.3.5. Other formats 47

3. PART THREE – Other information 48

3.3.1. Omission of publisher 49

3.3.2. Omission of publication date 49

3.3.3. Inclusion of author identifiers 49

3.3.4. Most recent data 49

3.3.5. File size 49

Further reading (bibliography) 50

Glossary 51

Annex 1 – Checklist (long) for footnotes or reference-list entries 53

Annex 2 – Checklist (short) for footnotes or reference-list entries 54

Annex 3 – Diagram for footnote and reference citation style with date format compliant with DCAT-AP 55

Annex 4 – Diagram for footnote and reference citation style with date format compliant with the *Interinstitutional Style Guide* 56

Annex 5 – Examples of citations with date format compliant with DCAT-AP 57

Annex 6 – Examples of citations with date format compliant with the *Interinstitutional Style Guide* 59

Foreword

Citation is the process of indicating what external sources have been used to create content. This is a guide to the citation of datasets.

At the time of writing in 2021, two issues dominate the headlines: (a) the spread of, containment of and recovery from the COVID-19 pandemic; and (b) the threat of, and required action against, climate change. These are classic political issues with different people and parties taking different views on the appropriate actions to take. Debate is vigorous and positions become entrenched. But they are both very modern issues in that they depend crucially on data.

The policies being proposed to deal with COVID-19 and with climate change are backed up by expert analysis. And the analysis builds on data produced by other experts. So why do proposed policies differ? Explaining that is not in the gift of this guide; however, as a minimum, observers should be able to determine whether the conclusions are being derived from the same data.

Different people can sensibly come to different conclusions after looking at the same data. But observers need to know what data has been used, where it came from and how it differs (if at all) from data used in other analyses.

This guide looks at the importance of providing links to this data in a robust and repeatable manner, and the ways in which this can be done so that it is widely recognised and understood. It covers both ensuring that data creators receive the credit due to them and deterring others from claiming the data as their own.

Following the guidelines presented here will ensure that the interests of all data creators, all data users and the general public are served when data is used.

The guide also touches on how best to make data citable, which looks at the problem 'from the other end of the telescope'.

Part One looks at the issues in data citation and makes general recommendations. Part Two contains specific formats and other elements of the 'recipe' for data citation. Part Three contains other useful information.

The busy reader might direct their attention to Part Two, but the earlier part of the guide contains material the recipe depends on, and this deserves their attention because it may affect the way they 'bake' their citations. In the annexes at the end of the document the reader can find checklists, diagrams and examples of footnotes or reference-list entries.

Data in the European Union

This guide was commissioned by the Publications Office (known as the OP) of the European Union (EU). The OP is charged with publishing and curating works created across the institutions, agencies and bodies of the EU. These include not only traditional text publications (such as reports and journals) but also websites and data services. The data services include data.europa.eu – the official portal for European data providing a single point of access to open data from international, EU, national, regional, local and geo-data portals.

Considerable energy and, consequently, public resources go into the creation and publication of data in the EU, and this guide aims to ensure that this data (and other data) is used responsibly and in ways that reflect and enhance this value for all stakeholders in the EU.

What is citation?

Citation is the process of adding information to a text or other material that indicates where an external resource has been used to create it.

The classic example of citation is a footnote in an original academic paper giving information about a different paper that was used to develop the material being written. This will typically include the name of the author and the name of the paper, together with where and when it was published – commonly in a scholarly journal. The reader of the original paper can then find the source material by going to a library and locating the cited paper. More recently it has become common to include a link in the citation that the user can click on and be directed to the cited material – either directly or through an individual or institutional subscription.

The original publication may not be an academic paper but perhaps a policy document or an internal report for local consumption within a company. The same factors apply: including information about the material used in generating the original allows a reader to find it if they want to.

It may be that readers do not themselves need to find the material that was used. It may be enough for them to find the name of the creator – which may give them confidence in its quality – or they may be satisfied to know that the original text builds on the same cited material as some other text and that they are therefore, in some sense, directly comparable.

Where data is involved, the principles are exactly the same. A newly written paper contains information about the data used in its creation so that a reader can find this data, or even (as above) just note who created it and draw some conclusions about its reliability.

This guide focuses on the citation of data and occasionally notes similarities with the citation of text publications. This should not suggest that there are no other entities that may need to be cited. A paper or report on fine art might need to cite a particular original painting (as opposed to a different version of the same subject by the same artist, a counterfeit by an unknown forger or a digital scan of the painting). The digital scan can probably be treated as data, as can dance notation, encoded music notation or embroidery patterns. But the performances and artefacts that result from them need separate, careful treatment in citation that is not covered here.

This guide also focuses on citation in the context of scholarly and policy documents, and makes the assumption that there is a considerable overlap of practice between these fields. Partly, this results from an (assumed) desire on the part of policy experts to adopt the principled rigour of scholarly research, but it also recognises common analytic processes. It should be understood that there is much activity outside these fields for which citation is just as critical. This includes documents produced in industry and commerce where data is used to support decision-making. The same recommendations should prove just as appropriate in those domains.

What is in this guide?

This guide contains recommendations for best practice, and it is hoped that data creators and users will be assisted by the advice. It also contains suggestions for circumstances where deviating somewhat from best practice may be acceptable. Finally, it contains warnings about deviations that may cause serious problems and examples of poor practice.

Recommendations look like this:

Recommendation: Users of this guide should read Part One as well as Part Two.

There are 'user stories' that give context to the value and importance of data citation. These look like this:

User story: An analyst was asked to provide a report on air quality. They cited the data they used and another analyst was able to use a different methodology with the same data to verify the results.

Finally, the guide contains examples of full citations or the elements that, together, make them up:

Example: https://doi.org/10.1000/182 (the DOI Handbook)

Where there is advice to avoid a practice, this appears as an example with a cautionary note:

Example to avoid: Foster, Dr G. (professional and academic titles should be omitted in names in citations)

Who is this guide for?

None of the recommendations presented here is carved in stone, but respecting them will make publishing and using data much more effective.

This guide does not in itself create any obligations to cite data in particular ways, or indeed at all. However, some of the recommendations include requirements that are contained in the *Interinstitutional Style Guide* (ISG) [1], which is published by the OP. These should be followed for works generated in or for the institutions of the EU. These are signposted by the phrase 'When subject to the ISG'. Other users are not

[1] http://publications.europa.eu/code/en/en-000100.htm

bound in this way and may be subject to other style guides. Information to assist them is signposted by the phrase 'When not subject to the ISG'.

Elsewhere, the recommendations can be seen as best practice when using data sourced from the EU institutions.

Others may take the recommendations at face value knowing that they are sanctioned within the EU institutions.

Acknowledgements

The OP.C.4 team and the author would like to thank the colleagues from Eurostat.B.4, JRC.I.2, JRC.I.3, JRC.I.4, OP.A.1, OP.A.4 and OP.B.4 for their contributions.

1. PART ONE – Citing data is important

1.1. Why cite?

The story goes that a university lecturer's 5-year-old son went into school with a project on farm animals – complete with footnotes saying where the pictures had been traced from. The teacher told the lecturer this wasn't really necessary for a 5-year-old. 'If you got more of your students to footnote their work, I'd be failing fewer of mine,' he responded.

In the world of writing, citation has long been considered important. It is regarded as part of the scholarly method and its techniques are fully embedded in publishing (and indeed in creations that are not intended for publication). For data, the importance of citation is increasing – partly due to the increasing importance of data in both scholarly and policy publication, but also due to the recognition of the value of citation and (to some extent at least) the emergence of tools that assist accurate citation. There is more information about these tools in Section 1.4.2.

In the cases of both text and data, the motivations for proper citation are similar. These are set out in this section.

1.1.1. Credit

The creation of datasets is not without cost. The creators make them available for reasons that include political mandates (see Section 1.5.2), scholarly principles, commercial benefit, ego and self-promotion. In any case, providing credit to the creator and the provider of data is part of a feedback loop that encourages further creation. In particular, providing a citation for a particular entity can inform the creator that it has proved useful and that further work in this area may be fruitful.

To an ever increasing degree, citation acts as a critical measure of success in academic and associated fields. Although citation of articles is seen as a measure of success, citation of data does not yet have the same kudos. The automated measurement of citations is inescapable, though it is, by some practitioners, regretted. Without the citation of datasets, these measures cannot be attributed to the appropriate creators. More recent measures of relevance in the form of 'altmetrics' aim to gauge the impact of publications (which include text and data) outside the formal structures of citation. This includes social media presence, and such measurement requires that many of the same formalities are observed, in particular the use of codified identifiers (see Section 1.2).

User story: A research team is investigating wildlife density in rural areas. They use data on animal food consumption from an earlier study, which they cite in their report. A separate research team uses the citation to retrieve the same data in their attempt to reproduce the experiment. The same findings are obtained, and the original conclusions are confirmed.

1.1.2. Transparency

Data citations lead the reader to the used data. But they also lead to the methodology used to capture or create it, which may be of considerable interest in evaluating how appropriate it is as source material for subsequent work.

Researchers try to reduce the biases that affect their work, but some inevitably remain. Good citation of the data used downstream allows others to review how any remaining biases affect the derived efforts.

More generally, access to the metadata describing the cited dataset enables deeper understanding of the scope and granularity of the source data, and consequently of the way it has been used.

1.1.3. Integrity

Just as the creators of resources used by others deserve credit, so too should the users of these resources not take that credit when it properly belongs to others. This is a major issue in academia, where failure to cite resources is seen as a form of plagiarism.

The citation of data is a clear indication to the user that it was generated by someone else and that it should not be confused with data generated directly.

User story: A postgraduate student retrieves and analyses data about chemical reaction rates and includes this data in a thesis. They fail to cite the data and the thesis gives the impression that it was collected by the researcher in person. The institution has an academic integrity policy that treats this as plagiarism, and the thesis is rejected.

1.1.4. Reproducibility

In science and other quantitative endeavours, a result is not considered definitive unless it can be reproduced independently. Where data from a third party forms part of the input into an experiment, it is critical that the reproduction of results uses precisely the same input as the original. The citation of the data used allows a subsequent experimenter to access this data and factor it into their attempts to reproduce the results originally obtained.

The same considerations apply in policy determination. Someone checking the analysis leading to a political decision will expect to be able to verify the conclusion by accessing the same data. Citation of this data in the initial analysis allows this to happen.

Similarly, publishing the results of studies in a way that allows them to be cited without ambiguity enables others to reproduce them. Additionally, publishing the data in a citable form enables analysis of differences observed in subsequent exercises. Both of these factors contribute to the robustness of the scientific and analytical process.

User story: A researcher fraudulently manipulates readings in an experiment. Institutional policies require them to publish the data in a suitable form. Other researchers use the citation to obtain the data relating to the readings and demonstrate statistically that it has been tampered with.

Comparability is a somewhat softer form of reproducibility. An experiment or analysis performed using a particular dataset becomes more useful if other different, but connected, exercises are performed using the same underlying data.

User story: A research team extends the work on wildlife density (above) to urban areas. They use the citation of food-consumption data in the earlier report to exploit the same data in the study so that the two sets of results are directly comparable.

1.1.5. Reuse

The availability of data for reuse is now a significant driver of data management in many fields.

By citing data used in work, a giant distributed catalogue of useful datasets is created. Readers can see what data was used and, referring to the citation, access it and use it for other purposes – even if they are entirely unconnected.

User story: An acoustics research team uses a published library of ambient sound recordings while studying audibility thresholds. They cite this library in their research report. A different research team is studying the impact of traffic noise on communities and finds the earlier report with its library citation. They are able to access and use the library in their work.

Similarly, making data citable improves its availability for reuse.

User story: The acoustics research team processes the sound recording library in a way that they show improves the sensitivity of tests that use it. They publish this derivative library in a citable form and, as it is used and cited, more users become aware of its availability and utility.

1.1.6. Text mining

Increasingly, textual materials are consumed not just by human readers but also by machines doing text analysis, or 'text mining' as it is sometimes called.

The intelligence of the machines that do this 'mining' is considerable: they can infer a lot about what they are analysing from its context. However, their job is made much easier if they know what to expect. This is one of the reasons that it is important to include citations in a form that is standardised and will be immediately understood as a machine attempts to process and interpret it.

The use of the standard formats set out in Part Two will aid not only humans but also machines as they seek to understand the dataset that is being specified in the citation.

1.2. Specifying the cited data

It is important that the citation has precision – that is that the cited data is accurately defined by the citation that references it. This is a link with two ends: the end that is rooted in the source document and the end that is attached to the target dataset. Both are critical.

1.2.1. Precision in the source

The citation must make clear exactly what data from the cited dataset has been used. In particular, when multiple datasets are combined in an analysis, the provenance of the different contributions and the ways they have been combined must be clear.

User story: A policy paper on improvements in domestic-appliance energy efficiency uses data from both official sources and a trade association. Each is properly cited, but it is not clear which source has been used in which calculation.

This can be avoided if the text is drafted to make the mapping between source and usage clear. This is not usually possible in the citation itself, but should be included elsewhere.

Recommendation: Where multiple data sources are used in conjunction with each other, the link between cited datasets and the use to which that data was put should be unambiguous. In particular, the processing used to generate derivative data from source data should be clear.

1.2.2. Specification of the target

The creator of a citation is faced with a dilemma if the data used is itself republished from another source. This is quite common, as data is aggregated in so-called observatories and scoreboards. Citation of the derivative data is 'correct' in that any changes introduced deliberately or accidentally were present in the data used. But citation of the original data may be more useful to a reader and to the creator of the original data.

Recommendation: When citing data that has been aggregated by a third party the formal citation should be of the aggregated data, but checks should be made when it is retrieved that the original source is clear. If the source is not clear, consideration should be given to including a reference to the original source in an appropriate way.

1.2.3. Precision in the target

It is also necessary to be able to specify the data used with precision. It is not enough to say that data was obtained from a particular data provider (such as Eurostat or Zenodo), because thousands or millions of data selections correspond to that description.

It is of course possible to describe the dataset using natural language.

Example: Eurostat publishes a table containing a dataset called 'ECU/EUR exchange rates versus national currencies'.

More precise than this, however, is using a formal scheme of naming so that everyone agrees what a particular citation means. This naming typically takes the form of short strings of characters. Such strings are called 'identifiers' and are rather common in everyday life. Cars have registration plates and vehicle identification numbers. People have passport numbers and identity card numbers. Books have barcodes that contain a special code [2] that works at the point-of-sale terminal. Library books may have individual barcodes that let you check them out (and be chased for fines if they are returned late).

In the field of data citation (and indeed numerous other fields where the precise, long-lasting specification of entities is required), these codes are called 'persistent identifiers' or 'PIDs'.

1.2.4. Characteristics of persistent identifiers

The three key features of PIDs are that:

(a) they are managed to enable persistence (see Section 1.7.6);

(b) they allow users some assurance that any given identifier really is associated with the correct entity; and

(c) they are actionable, which is to say they can be sent to an online service that will return something more useful than the identifier string.

[2] The International Standard Book Number (ISO 2108) is a 13-character code that identifies books and occupies a subset of the Global Trade Item Number space so that it can be used in point-of-sale systems.

Typically, a PID provides direct access to the entity itself (or the means to get it via, for instance, a 'landing page') or to metadata about it.

In practice, the second and third characteristics are both delivered by an online resolution service – typically a web-based service that accepts identifier strings and returns a block of data about the identified entity. This may include the web location where the entity can be found. Thus, by 'resolving' an identifier string to its metadata, the assurance of identity can be obtained. Additionally, the entity itself may be obtained by following the link contained in the metadata.

The rest of the metadata in the block is critical for indexing and discovery, and PIDs are most useful when they are registered in systems.

In order to make an identifier actionable it usually needs to be expressed as some form of Uniform Resource Identifier (URI). URIs are the standard internet mechanism for identifying logical and physical resources and are recognisable by starting with the name of a 'scheme' and a colon. The most familiar type is the http or https URI, often known as a Uniform Resource Locator (URL) and acting as a web address.

1.2.5. Types of persistent identifier

There are numerous families of public PID competing for attention. Some are specific to particular types of identified entity.

Example: The International Standard Name Identifier (ISNI) identifies public identities of parties involve in creating content, while the Open Researcher and Contributor ID (ORCID) identifies authors and contributors in scholarly communication.

Other families of PID are used to identify arbitrary entities (digital resources, physical objects and even abstract concepts). This guide will note two of the most commonly encountered schemes: the Handle System and the Digital Object Identifier (DOI) system.

1.2.6. The Handle System

The Handle System [3] is a generic mechanism for associating an identifier string with a thing, through a resolution service that accepts the string and gives back a block of data that describes (through metadata) the thing and, if appropriate, a link to the thing itself.

The syntax of a Handle is rather simple:

prefix/suffix

It is possible to resolve a Handle in this form using its own specialised protocol, but a Handle is usually resolved using a web proxy server, and consequently it is often written using the web address of this proxy.

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

Example: 20.1000/100 is a Handle identifying a licence document. It can be written as a URI https://hdl.handle.net/20.1000/100 and clicking on this link takes you to the document.

Handles are used in many different areas, but are commonly used to identify data.

1.2.7. The Digital Object Identifier system

The DOI system [4] builds on the Handle System by adding a layer of governance, both technical and social. The focus of the DOI as a PID scheme is on persistence. There are mutual promises between the participating registration agencies that actually register DOI names to continue to maintain them if the issuing agency is unable to do so.

DOI is widely used in the scholarly world for journal articles and datasets. It has also found application in the movie/TV space for audiovisual assets and associated entities, and other uses are emerging.

The registration agencies include DataCite [5], which specialises in DOIs for datasets and specifies a metadata schema to standardise their description. Crossref [6] also acts as a registration agency, and deals with data as well as publications. Other DOI registration agencies also register datasets, and all collaborate on standards.

Currently, all DOI names have a prefix that starts with '10.', but this is under review and other prefixes may be seen. There is a specialist web proxy server that allows for the resolution of DOI names.

Example: 10.5281/zenodo.5378074 is a DOI name identifying a dataset of air traffic during the COVID-19 pandemic, and can be written as https://doi.org/10.5281/zenodo.5378074 to be actionable.

1.2.8. The Publications Office of the European Union and persistent identifiers

The Publications Office of the European Union (OP) is a registration agency within the DOI system and supports the assignment of DOI names to relevant publications and other types of content within the EU institutions.

As well as the generic web proxy server run by The DOI Foundation which is accessed by prepending https://doi.org/ to the DOI name, there is an EU-specific web proxy run by the OP which works by prepending https://data.europa.eu/doi/.

[4] https://www.doi.org

[5] https://datacite.org

[6] https://www.crossref.org

Example: 10.2766/960 is the DOI name for an EU report on education and training. It can be accessed through the landing page that is pointed to by https://data.europa.eu/doi/10.2766/960

As well as assigning DOIs to monographs, OP also now assigns DOIs to articles in journals, as well as datasets. By working in partnership with other DOI registration agencies, notably Crossref and DataCite, it is able to offer developed services for these content types to its clients, over and above a 'simple' resolution service.

1.2.9. Custom identifiers

In some cases, a dataset provider will create and maintain its own system for identifying its products. Being professionally developed, these are often as well managed as systems using standardised PIDs. Where they come from an institution with a sustainable operational model and a commitment to data availability in the long term, they may be as persistent as any other system. Users should satisfy themselves with respect to this persistence.

Recommendation: When a data provider offers a stable naming system for its datasets, these should be treated as PIDs and, if possible, written as URIs.

Eurostat provides framework for persistent identification through a structured naming system for its datasets. Its assurance of the persistence (see Section 1.7.6) not only of the data but also of the resolution system that allows it to be accessed by its identifier, creates confidence in the system.

Example: Eurostat monthly data on production in the construction sector is contained in a dataset called 'sts_copr_m'.

These Eurostat codes can be constructed as actionable URIs in various ways to address different needs. For citation purposes, a form that resolves to a landing page is typically most appropriate.

Example: The 'sts_copr_m' dataset can be constructed as a URI as 'https://ec.europa.eu/eurostat/web/products-datasets/-/sts_copr_m', which links to a landing page that provides metadata about the dataset and links to download it.

Alternatively, it may be appropriate to link to a view on the data itself.

Example: The 'sts_copr_m' dataset can also be constructed as a URI as 'https://ec.europa.eu/eurostat/databrowser/view/sts_copr_m/default/table?lang=en', which links to a default view of the dataset.

1.2.10. Using available persistent identifiers

When a PID is available for use in a citation, failing to include it will lead to avoidable ambiguity and uncertainty.

Recommendation: When a dataset has a recognised public persistent identifier, that identifier should always be included when the data is cited, preferably as an actionable link. This identifier should be used even if the location of the data or a local name is available.

1.2.11. Short-form persistent identifiers

Services are available that create URI identifiers that are still unique but shorter than the original. The guarantees of persistence some of these services offer are not always comparable to those offered by PIDs themselves.

For example, the shortDOI service is under the same governance as DOI itself and generates a shorter code that will have the same persistence characteristics and resolves to exactly the same information. However, a generic URL shortener will generate codes that are only as persistent as its operator.

Although an original identifier may resolve in exactly the same way as the shortened version, they are textually different, and two citations to the same data may appear completely different. This makes the use problematic unless the creator of the data has itself generated the short identifier.

Recommendation: A dataset citation should not use a shortened PID unless it was both generated by the creator of the data and recommended for citation use, for instance in 'cite this dataset as' information.

1.2.12. Local identification schemes

In addition to the public PID schemes above, it is common for publishers to create a local naming scheme for datasets. Though they rely on the local governance [7] of the scheme, it is almost always better to cite using this local identification scheme than to omit it. Where the dataset is available under multiple names, which may represent multiple locations, a careful choice of the name to cite should be made.

Recommendation: When a dataset lacks a recognised public persistent identifier but has been assigned a local name (typically a URI) in a structured scheme, that name should be used in a citation.

[7] Although users typically have little influence over this governance, it is critical in that it determines whether the naming scheme will be useful over time, and in particular whether the URIs continue to resolve correctly. Providers of such schemes need to take these responsibilities very seriously.

1.2.13. Link rot and broken links

One of the problems that PIDs are intended to solve is that web links are in practice extremely fragile and often do not work as they were intended to – become 'broken' – quite soon after they were created. People change the architecture of websites so that resources (including datasets) are to be found at a different location, and in extreme cases organisations change their name and migrate to an entirely new domain. This process by which a once functional link ceases to work is called 'link rot'.

A parallel problem is 'content drift', where data is silently changed or replaced. This is a governance issue for PID systems. It can be appropriate for data identified by a PID to change, but this must be explicit, and these identifiers are not in general suitable for use on their own in data citations.

PIDs do not provide a 'magic bullet' to solve the problem, but they do provide tools that allow links to be made to continue to work in the medium to long term. If an organisation reorganises its website or changes its domain name it can register the new location of the dataset against the persistent identifier, and users can continue to access it by reference to its identifier rather than its original location.

Recommendation: When a dataset is generated and made available to others, it should, if possible, be assigned a persistent identifier to enable its accurate citation.

1.2.14. Other persistent identifiers in citations

The citation for a dataset typically includes the names of the authors and the name of the publisher. The question therefore arises of whether the citation should include any PIDs identifying these parties. There are several identifiers that identify parties: ORCID ([8]) identifies researchers and contributors in the scholarly domain (currently strongest in science, technology and medicine); ROR ([9]) identifies organisations such as university departments and libraries; ISNI ([10]) identifies the names of both individuals and organisations involved in creating things.

It is not currently customary to include any of these identifiers in the citation, and the argument against this is that the purpose of including authors' names is not primarily to identify those parties but to identify unambiguously the entity being cited. The main tool for this is a PID for the cited entity, and adding PIDs for the authors often helps hardly at all.

The metadata associated with the PID for the cited dataset (which is hopefully accessible by resolving it) should contain the authoritative information about the creators and publisher (including their PIDs), so it is not necessary or appropriate to include it in the citation.

However, ambiguity sometimes arises where no PID for a dataset is available and a name is common – 'J. Smith' or 'Trinity College' for instance. In these circumstances, including an identifier in association with a name would be appropriate.

([8]) The Open Researcher and Contributor ID (ORCID) has a website at https://orcid.org

([9]) The Research Organisation Registry (ROR) has a website at https://ror.org

([10]) The International Standard Name Identifier (ISNI) is specified in ISO 27729 and has a website at https://isni.org

Recommendation: PIDs for authors and publishers should not normally appear in a citation unless it resolves ambiguity.

Example to avoid: Berners-Lee, T. https://orcid.org/0000-0003-1279-3709

1.3. Data that changes

Along with fixed datasets that are created and never change, there are two types of datasets where the data varies over time.

- **Continuing datasets.** Data that is from time to time added to, but where older data is never changed once made available.
- **Dynamic datasets.** Data that changes even once made available.

Examples:
Fixed dataset – table of results from an experiment.
Continuing dataset – table of website accesses per month with new row added every month.
Dynamic dataset – best estimates of monthly national retail spending where old data is updated as better source information becomes available.

Citing a fixed dataset is reasonably straightforward. Citing a continuing dataset requires care. Citing dynamic data is complex and may involve compromises. The issues are summarised in this section.

1.3.1. Continuing datasets

When new data is added to a dataset containing time-series data, the addition is usually recent data that has just become available. That is the focus here, but sometimes the new data relates to an extension of the scope of the dataset. Care should be taken to ensure that the citation accurately reflects the data actually used. Guidance on cases where data changes other than by the addition of new, recent information is in Section 1.3.2.

Where a time-series continuing dataset has more data added from time to time it is helpful to include in the citation the date on which the relevant data was obtained. This is done by including the date of access in the citation. By including this date of access, a subsequent user may be able to determine which elements of the dataset were used and which have been added since it was accessed.

Recommendation: Where a continuing dataset is cited, the date of access should be included in the citation if this is sufficient to define the data used.

Sometimes it will be enough to give just a date. In other cases, it may be important to be more precise and the time and time zone should be included as well. Section 2.3 has information on the way to do this.

Examples:
accessed 2 July 2020
accessed 2020-07-02T15:36:17Z

Citing an access date can risk being misleading if the dataset does not include the date on which each element was added, particularly if the elements relate to date-related information.

User story: A dataset containing pollution levels contains the results of averaging samples over each calendar month. The sampling, processing, analysis and checking takes a few months, and this time varies when the period includes public holidays, vacation seasons, etc. When data is added, the date of addition is not included. A citation includes an access date. A user is unable to determine what data had been included in the dataset at that date.

The date of the update is less important (for reproducibility in particular) than the date of the most recent data present in a continuing dataset. A version number may provide more certainty.

This problem can be avoided by including in the citation the date associated with the most recent data available on the date on which it was accessed. This can ensure that later data that becomes available after the citation is created is not included in it.

Recommendation: When a continuing dataset of time-series data is cited, consideration should be given to specifying not only the access date but also the date associated with the most recent data included when the data was accessed.

Recommendation: Use of the term 'most recent data' should be considered when citing a time-series continuing dataset if there is any possibility of ambiguity about what data is included in a continuing dataset being cited. This should reference the date associated with the most recent data available in the dataset when it was accessed in the activity for which is it being cited. The precision of the information in the 'most recent data' citation should match the data frequency of the dataset.

Example 1: A dataset contains monthly data on traffic levels. When accessed in July 2020 the most recent available data was for May 2020. The citation would include: accessed 2020-07-02, most recent data 2020-05

Example 2: Another dataset contains daily traffic data, but it is updated monthly for full months. The citation here would include: accessed 2020-07-02, most recent data 2020-05-31

1.3.2. Dynamic datasets

Where data may be changed after initial publication, the situation is much more complex. Including the date of access removes some ambiguity, but may not be adequate from the perspective of repeatability: some data may have changed between the point of citation and the point of reuse.

1.3.2.1. Snapshots

The dataset publisher may provide 'snapshots'. Snapshots are copies of a dataset taken at specific points in time and not updated. Using and citing these snapshots avoids the problem of ambiguity.

Recommendation: Where dynamic data is provided with periodic snapshots, the snapshot used should be cited in the usual way as a 'version' of the dataset.

Alternatively, when this makes sense and the licence conditions permit (see Section 1.5), the user of the data can take their own snapshot of the data and publish it. All the considerations in Section 1.5 about publishing data apply, and this should not be undertaken lightly. It does ensure that any future reuse of the data will be undertaken from precisely the same basis, however.

This recommendation mirrors best practice in citing websites that have dynamic content ([11]).

Recommendation: When dynamic data is provided without snapshots, consideration should be given to taking a snapshot of the data as used and publishing it for the reference of future users.

Recommendation: When a snapshot of a dataset is taken by the user, its citation details should be included in the original citation as an extension of the access information. This should appear as 'snapshot available as' followed by an actionable (clickable) PID or 'snapshot available at' followed by an http URI (web address).

Example:
accessed 2020-07-02 snapshot available as http://doi.org/10.5281/zenodo.999999

1.3.2.2. Citing queries

It is possible to address the problem of dynamic data in a much more sophisticated manner, but this requires the dataset to be made available in a specific way. A particular approach was developed by the Working Group on Data Citation of the Research Data Alliance (RDA), and is recommended by them.

The details of this approach are beyond the scope of this guide, but involve the data provider applying timestamps to each data element and revisions of them. The citation then takes the form of a query made to the database. This places obligations not

[11] For example, ISO 690 – 'Information and documentation – Guidelines for bibliographic references and citations to information resources' contains the following in clause 8.14.6: 'If the cited Web page is dynamic and a specific version is used, an archived version of the resource shall be cited. If an accurate archival copy of the page is not found, one shall be made.'

only on the creator but also on the data repository to include timestamps and enable a compatible query function. The combination of timestamp and query ensures that a user encountering the citation will be able to execute the query and obtain exactly the same results as were obtained originally. Reference to the RDA documentation ([12]) is recommended to understand this procedure.

In existing systems, Eurostat allows the creation of a 'bookmark', which represents a query into their database. Such bookmarks can be expected to be persistent, and careful reference to the Eurostat documentation is essential when using these bookmarks in citations, where they should be accompanied by other identifying information. The bookmarks act as a PID, or at least as a component of a PID that identifies the result of the query.

1.3.2.3. Versioning and citation

Where the update rate of a dataset is manageable, versioning may be appropriate. This assigns a new PID to a dataset every time it changes. Obviously, it is not appropriate for truly dynamic datasets. Some repositories encourage the assignment of a new identifier (typically a DOI name) with every update to the dataset. Citing this identifier removes all ambiguity. They may also allow the assignment of a 'concept identifier' (typically called a 'concept DOI'), which acts as an identifier of the group of all the connected datasets. Although it identifies the group, it usually links, on resolution, to the most recent version.

A concept PID should not normally appear in a citation because the data it is linked to will likely change over time as new versions are created. It therefore fails the reproducibility and reuse tests mentioned in Section 1.1. A citation should normally use the appropriate PID for the specific dataset that is being cited.

However, there are occasions where the purpose of citation is more general than this and focuses on the creation, existence and ownership of a dataset, possibly giving credit for its publication. In those circumstances, citing the concept PID is appropriate.

User story: A policy paper reviews funding for projects generating a class of data resources. It notes several continuing and dynamic datasets in the class. It cites these datasets by reference to their concept DOIs.

[12] The approach is summarised at https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf, and further information is contained within the archives of the Research Data Alliance at https://rd-alliance.org

1.4. Creating good citations

1.4.1. Authoring tools

Most software used for creating documents (word processors, desktop publishing systems, website authoring applications) will have facilities for the automatic creation and formatting of citations. While these systems will not create or check that citations are correct, they will ensure that, for example, footnotes appear at the bottom of the correct page and bibliography entries appear at the end of the document or the chapter, depending on the settings selected. Importantly, they will also deal with very long footnotes that need to be split across pages so they do not occupy more space than the core text, and move footnotes between pages if citations are so concentrated that a large number would dominate a single page.

The precise capability of particular software packages is beyond the scope of this guide but reference to manuals and online training materials is recommended.

1.4.2. Citation management tools

Beyond word processing software, there are citation management tools that work alongside it, or with desktop publishing or (in particular) typesetting software. These effectively build a database of citations, possibly pulling the accurate information direct from the cited source, and allow the addition of the citations into the finished document.

The popularity of these tools varies over time, but BibTeX ([13]) has been stable for many years and has spawned numerous additional tools that work with it. The commercial software and service Mendeley ([14]) has also been under active development for some years. Many others are available.

These systems have a steep learning curve, but they can save time, particularly where numerous datasets are to be cited in multiple documents.

Again, the detailed operation of these tools is beyond the scope of this guide, but it should be noted that some of them have specific features aimed at the citation of datasets.

([13]) http://www.bibtex.org

([14]) https://www.mendeley.com

1.5. Rights and licences

1.5.1. Licences and citation

The licence under which data is made available is not really a citation issue. However, there are a number of interactions between the licence governing the use of the data and its citation. Certainly, the licence governing the data should be accessible via the citation (for instance by resolving the PID), even if it is not contained within it.

1.5.2. Data published by EU institutions, bodies and agencies

It is becoming recognised within the European Union that there is considerable benefit and value in making data that is generated within the institutions as widely available as possible. This has been reflected in a series of policy changes that now mean that much official data can be reused.

The key initial regulation ([15]) and decision ([16]) focused on the policy that documents within the European Commission should be made available for reuse. In 2019, a study ([17]) on available licences was published and the European Commission decided ([18]) that documents should, by default, be made available under a Creative Commons licence requiring attribution.

That decision adopted the Creative Commons Attribution 4.0 International Public License (CC BY 4.0) as an 'open licence for the Commission's reuse policy' under the earlier decision. Although reference should be made to the licence itself, this broadly allows copying, reuse and the creation of derivative works, providing only that the source of the document is declared in an attribution statement. In particular, the licence allows commercial reuse.

Some documents are not available under this licence for reasons including personal privacy, security, public interest and the inclusion of material covered by third-party rights. Documents are marked with the relevant licence.

In parallel with progress on documents, policy on the reuse of data has evolved. The same 2019 decision recognises the benefit of releasing raw data, metadata and 'other documents of comparable nature' to the public domain and allows the use of the Creative Commons Universal Public Domain Dedication deed (CC0 1.0). This instrument goes to great lengths to ensure that the information released under it is available for use and reuse without any restrictions. Again, there are exceptions similar to those for

[15] Regulation (EC) No 1049/2001 of the European Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents (https://eur-lex.europa.eu/legal-content/en/ALL/?uri=celex:32001R1049).

[16] Commission Decision of 12 December 2011 on the reuse of Commission documents (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX %3A32011D0833).

[17] Central IP Service of the European Commission, *Reuse Policy – A study on available reuse implementing instruments and licensing considerations,* EUR 29685 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-00670-1 (online), doi:10.2760/95373 (online), JRC115947.

[18] Commission Decision of 22.2.2019 adopting Creative Commons as an open licence under the European Commission's reuse policy.

documents, and there is marking of datasets to indicate their licensing status – typically on the landing page from which they can be downloaded.

The availability of data under these licences is intended to stimulate economic benefit, and the value of this data is increased in use if it is cited correctly so that its provenance is clear, and its further use is promoted. It should be recognised that the extremely permissive CC0 instrument does not devalue the information, and citation remains a critical factor for all the reasons set out here.

Recommendation: Data should be properly cited even if it is in the public domain (CC0).

1.5.3. Unpublished data

It may happen that a dataset that is used is not generally available or accessible to others. It may be proprietary or so granular that it cannot be made generally available without identifying individual persons or bodies. It is common for summary data to be made freely available while the granular information is restricted. For example, Eurostat makes 'microdata' [19] available only on restrictive terms that protect the interests of the data subjects.

Such data should still be cited in the same way as other datasets. This is for the two reasons described below.

Firstly, the dataset may become available later because the owner or controller of the data changes the access terms. This is quite common, as proprietary data with a high market value is released without charge once its market relevance has passed. Alternatively, the data may be released because of legal proceedings or similar. Citing the data while it is not accessible means that a user can correctly obtain access to the data if it is later released.

Secondly, the private data may have been obtained by another user under special terms. Provided both users cite the data in the same way, readers of their reports can be sure that each analysis is based on the same underlying information.

The same considerations apply if the data is available only on subscription terms. It should be cited in the usual way for both of the reasons above.

Unpublished datasets may not have robust identification schemes or assigned PIDs, so close attention should be paid to specifying the cited data accurately.

Recommendation: Data should be properly cited even if it is not generally available, not fully available or available only on restrictive terms.

[19] Eurostat publishes information on microdata at https://ec.europa.eu/eurostat/web/microdata/overview

1.5.4. Informing the user about reuse access terms

As noted above, licensing and citation are formally separate. Although the terms under which data is available are important information to the user or reuser of such data, they should not be included in a formal citation. However, if they are considered important, they should be communicated alongside the citation.

Recommendation: Terms for the reuse of cited data should not be included in a formal citation. These terms may be noted inline in text or as a separate note or footnote.

1.5.5. Meeting the CC BY requirement

It was noted above that information provided by the European Commission will often be available under the CC BY 4.0 licence. Other information may be provided under the same licence (or very similar licences such as earlier versions of CC BY).

These licences require attribution of the licensor, and the question sometimes arises as to whether the citation of the data meets that requirement. This guide does not offer legal advice, but highlights two issues that the user might consider. The first consideration is whether the user has done anything with the data that implicates one of the acts controlled by copyright (reproduction, distribution, etc.), and the second is whether the information in the citation satisfies the requirements written into the licence.

It should further be noted that these licences require an indication of whether the licensed matter has been changed. The user must interpret this requirement carefully – but this issue relates to the republication of the data rather than its exploitation in an analysis. Whether or not the matter has been altered is probably not appropriate for inclusion in a citation that is intended to signal the inputs into a process rather than the process itself.

1.5.6. Privacy and the general data protection regulation

Some datasets contain personal data and are therefore potentially subject to privacy legislation such as the general data protection regulation (GDPR). GDPR compliance is important but outside the scope of citation. However, there are issues that need to be handled carefully.

A dataset that contains personal data may be made available on the condition that any analysis that depends on it is published only if the personal data is anonymised. Unless care is taken, it is possible for the citation of the data to unintentionally expose un-anonymised details.

User story: A medical dataset is analysed to isolate individuals with a certain trait and investigate other factors that may be associated with it. If the number of individuals is small, it is possible that the citation could allow them to be identified even if the text is careful to prevent this.

Recommendation: When datasets contain personal data, care should be taken to ensure that the citation of the dataset does not allow the identification of the individuals concerned.

1.6. Data mining and data harvesting

In some cases, the data used in a project may come from a large number of sources. As more data becomes openly available on the internet, and as this data becomes self-describing, processes to 'mine' and 'harvest' data will become more common. Autonomous software agents will seek and retrieve data from many sources that make it available with appropriate licensing.

It may not be realistic to cite all the sources. Indeed, the identity of all the datasets used may not be known to someone creating a citation. In these circumstances it is not possible to offer a firm recommendation on what form of citation to use, but the user should try to create something that achieves the objectives set out in Section 1.1.

Recommendation: When numerous data sources have been accessed in a fragmentary manner, consideration should be given to how they can be cited to provide most effective guidance on the sources used.

1.7. Publishing data for good citation

This guide is about citing data, but recommendations about making data available are relevant. If done well, it will be straightforward for users to cite the data and meet the objectives set out in Section 1.1. This will mean that the data is accessible to those encountering the citation, that the citation offers precision in specifying the data that is cited and that the data and its citation have persistence, thereby ensuring they remain available to those who wish to secure access.

A discussion of the use of published data for other purposes is contained in Section 1.1.5, but we focus here on the steps that can be taken to make citation easier and more effective.

The recent trend in the management of datasets is that they should, as far as possible, conform to the 'FAIR' principles ([20]). These require that data should be findable, accessible, interoperable and reusable. Many of the recommendations here contribute to these aims (the details of which are beyond the scope of this guide, but include important considerations for the publication of datasets).

The European Commission also addressed this issue in its strategy for data ([21]), where it noted the FAIR principles and the importance of all of the factors in them. It set out ways to overcome obstacles to their adoption in the data economy

[20] There is a comprehensive explanation of the FAIR principles at https://www.go-fair.org/fair-principles/

[21] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – A European strategy for data, COM(2020) 66 (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066).

1.7.1. Availability

For the cited data to be useful, it must be made available to those encountering the citation so that they can get a copy of it. This effectively means that the data must be accessible over the internet – the days of requesting a copy of the data on a magnetic tape or disc are behind us (except perhaps for the most enormous datasets or for users without access to high-bandwidth connections).

1.7.2. Local hosting

The data may be hosted by the creator themselves on a website or FTP site, though careful consideration is required before committing to this approach. The amount of traffic generated by such data initially is probably small, but can become very large if events suddenly create interest in the dataset. This can cause an interruption in the availability not only of the hosted datasets, but also of other services that share the same bandwidth or are hosted on the same machines.

Recommendation: Before cited data is hosted on a machine under the control of the creator, protection of that system in the event of sudden increases in download volumes should be considered.

The local hosting of data should also be understood as a long-term commitment. Data citations may continue to be important for years, decades or longer, and the creator should ensure that the data itself will be available for at least as long. They should understand that their association with the organisation hosting the data will likely have long ceased by this time.

Recommendation: Before cited data is hosted on a machine under the control of the creator, consideration should be given to the ability and willingness of the hosting organisation to continue to do this in the long term and after the creator is no longer connected with it – at least as long as the cited data remains relevant.

1.7.3. Data repository

As an alternative to hosting it themselves, the creator may make arrangements for the cited data to be hosted at a generic data repository such as Zenodo ([22]). Although there are no guarantees about the continued support for such services, the aggregation of many datasets that have value means that there is considerable political pressure for the data to remain available in the long run.

Recommendation: Before using a generic data repository, creators should ensure that the governance and sustainability of the repository mean that it is likely to remain in operation, and to continue to support the cited data for at least as long as the data remains relevant.

([22]) https://zenodo.org

1.7.4. Precision

As described in Section 1.2, the use of a persistent identifier to specify a dataset with precision is helpful.

If a generic data repository is used to hold the dataset, a persistent identifier will often be assigned by the operator. This can have a very beneficial effect, provided the identifier relates exactly to the data that the creator wishes to have cited.

User story: A dataset containing pollution measures across numerous cities is uploaded to a repository and a persistent identifier is assigned to it. The data a user wishes to cite relates to only a single city, and therefore the assigned identifier is inappropriate.

In these circumstances, the creator will either need to upload and register the smaller dataset separately, or work with the repository provider to assign an identifier to the subset. The Research Data Alliance recommends that this be treated in the same way as dynamic datasets (see Section 1.3.2), and that a query that generates the subset be included in the citation.

Recommendation: A dataset made available because it is to be cited should whenever possible be assigned an appropriate persistent identifier that accurately specifies the data that has been cited.

In parallel with assigning a PID, the creator of a dataset needs to document the dataset so that it can be understood by those trying to access it. This metadata needs to accompany the PID registration.

Recommendation: When a dataset has a PID assigned, comprehensive metadata for the dataset should be registered with it to allow discovery and reuse. This registration should follow standard practices wherever possible and use common schemas that are recognised by the community.

1.7.5. Getting a PID for a dataset

Those creating datasets within the context of the European Union have a number of options for obtaining a PID for a dataset.

The OP manages the data.europa.eu portal, which assigns PIDs to all datasets from EU institutions and bodies, and DOIs when requested.

The Joint Research Centre of the European Commission has its own persistent URI scheme to generate PIDs. It is planning its own repository, which will be able to assign DOIs through the OP.

At the simplest level, the Zenodo (23) repository can be used to store the data and to assign a DOI to it. Zenodo is hosted by CERN (as a 'marginal activity' (24)) and partially funded by the EU.

1.7.6. Persistence

Sections 1.7.2 and 1.7.3 make recommendations about the longevity of the hosting arrangements for data that has been cited. This is one aspect of persistence, but not the only one. Data curation expert Andrew Treloar has pointed out (25) that equally important are the persistence of the identifier, the persistence of the binding (linking) between the data and its identifier, the persistence of the service that takes you from identifier to data (resolution) and the persistence of the service that lets you make updates to the resolution service. All of these are important if a persistent identifier plays a central role in citation.

The examples presented in this guide use identifiers that are generally regarded as likely to remain persistent, but users need to be aware of the other factors above that may affect the value of a citation over time.

Recommendation: When a dataset is made available with an associated persistent identifier that enables precision in citation, consideration should be given to the persistence of all aspects of the identification system.

1.7.7. 'Cite this dataset as'

Finally, when a dataset is made available in circumstances where it is expected that people will want to cite it, it is very helpful to these users of the data for the dataset to be associated with information about how the dataset should be cited. This can provide all the information that is outlined in Section 2.1. The information can be transferred to new publications without fear of corruption and, importantly, can be copied directly into citation management software as mentioned in Section 1.4.1.

Recommendation: Repositories that make datasets available should provide citation information that can be reused by parties wishing to cite the data.

(23) https://zenodo.org

(24) https://about.zenodo.org/infrastructure/

(25) https://andrew.treloar.net/research/diagrams/five_persistences.pdf

2. PART TWO – How to cite data

Part Two contains the technical reference material that an author needs to create citations that meet the recommendations of this guide.

It covers the components – who the author is, what the data represents and so on, where in the author's document they go and how to arrange them for different purposes.

2.1. The components of a data citation

This section summarises the 'ingredients' of a citation, and Section 2.2 will demonstrate how to assemble them for different styles and purposes.

Certain fields are common to all the styles. The careful inclusion of the correct content for these fields will ensure accurate and useful citation.

This section provides technical details and outlines special circumstances in the specification of the components of a citation. Because these are all 'recommendations', the special formatting of recommendations in boxes has been omitted here.

2.1.1. Authors

The authors were responsible for creating the dataset. They include organisations ('corporate authors') in whose name the dataset is made and the personal authors who did the work. In some cases, the organisation will be noted and the workers remain anonymous; in others the authors get the credit and their employer or hosting body steps back. Sometimes both have their names displayed. The metadata of the dataset will show which option the creators have chosen, and that should, in general, be followed.

When the metadata associated with a dataset is examined, the parties termed 'authors' here may be referred to as 'contributors' or using similar terms.

There is a trend to include many author names in academic work-product, particularly in science when large teams are involved. It may be unrealistic to include all such names in a citation involving many authors, and it is usually unnecessary for precision.

Where there are three or fewer authors, including any corporate author, they are normally listed individually, separated by commas, except that the last author is preceded by 'and'.

Example:
Newton, I. and Hooke, R.

Where there are more than three authors, only the first three (including any corporate author) are normally cited, separated by commas and followed by 'et al.' [26].

Example:
Einstein, A., Bohr, N., Fermi, E. et al.

However, if space is restricted or the first named author is considerably more notable than the others, then citation can be restricted to the first author followed by 'et al.'

Example:
Einstein, A. et al.

Alternatively, if space is not limited, more than three authors may be included. Consideration should be given to how useful this is: the purpose of citation is to ensure the cited dataset is precisely recorded, and the persistent identifier (see Section 2.1.9) is the most reliable mechanism for this. Resolving the identifier will reveal the full list of authors, who thereby receive due credit.

Example:
Einstein, A., Bohr, N., Fermi, E., Planck, M., Schrödinger, E. and Hawking, S.

2.1.2. Corporate author

The corporate author is the institution or corporation that is responsible for the creation of the dataset. It might not be the publisher of the dataset. In some cases there will not be a corporate author, and the personal authors will have acted on their own behalf.

Where the corporate author is an institution of the European Union, the following special rules apply.

For institutions acting as a whole, the full name is used.

Examples:
Council of the European Union
European Commission
Court of Justice of the European Union

For parts of institutions, the institution and its component are separated by a comma.

Examples:
European Commission, Joint Research Centre
European Commission, Directorate-General for Research and Innovation
Council of the European Union, General Secretariat of the Council

[26] This expression abbreviates the Latin phrase 'et alii' ('and others'), and is the equivalent for people of 'etc.' ('et cetera'), which is used for things.

Finally, interinstitutional bodies are named in full.

Example:
Publications Office of the European Union

2.1.3. Personal author(s)

The personal authors are the natural persons who have created the dataset and are making it available for publication.

As noted above, where the dataset is published in the name of the corporate author, the names of the personal authors might not be disclosed, and they do not need to be included.

Names are written with the surname/family name first. It is separated by a comma from the initial (or initials), with full stops, or alternatively the full first name.

Examples:
Einstein, A.
Einstein, Albert

Honours and qualifications should not normally appear in a citation unless they are required to remove ambiguity. Similarly, the year of birth or death (frequently used in bibliographic contexts) should not appear unless it is required in order to remove such ambiguity, which may be more common in some cultures.

Authors' names should appear as they appear in the metadata of the dataset, particularly with respect to spelling, capitalisation and hyphenation.

The same convention of family name first should be carefully applied to names normally written with the family name first anyway. Where someone is known by both a given name in their first language and an adopted first name in another language, the attribution associated with the original publication of the dataset should be followed.

2.1.4. Dataset title

The dataset title is the name assigned to the dataset by its creators, under which it is published and by which it will be recognised. Where a dataset is known by both a full name and a short code, the short code should follow the name in parentheses.

Example: Eurostat data on pensions has the title
'Pensions in national accounts' (nasa_10_pens)

Note that the short code is outside the quotation marks.

When subject to ISG the title can be cited in quotation marks as in the example above, or in italics. When the title is presented in italics and is in the English language, all the main words should be capitalised.

Example:
Pensions in National Accounts

2.1.5. Version/edition

The version or edition of the dataset, as specified by the authors. This is typically used with continuing datasets (see Section 1.3.1) and may be omitted if not stated or not available.

When present, the version number is prefixed by 'version' or the equivalent in the language used by the publisher.

Example
version 2.0.4

2.1.6. Publisher

The publisher is the entity that makes the dataset available.

The publisher should not be confused with the corporate author. Where the publisher is the same as the corporate author, the publisher may be omitted, with the single instance of the corporate author communicating all the useful information.

Care should be taken in identifying the roles of a publisher and a hosting platform providing access to the data. A publisher takes active decisions about what content it should make available and is involved in editorial decisions about how it is presented, marketed, etc. A platform may have rules about the type of content permitted, but it is passive and allows broad access to creators to make content available. However, it usually also hosts the metadata noted in Section 1.7.4, and plays an important role in making data discoverable.

For the purposes of data citation, it may be appropriate to regard the platform as the publisher, but care should be taken to avoid confusion where the publisher and the platform are distinct. The objective should be to make the citation precise to allow future access to the data used.

Example to avoid: Zenodo might be cited as a publisher when it acts as a repository if no conventional publisher is involved, The Zenodo-assigned DOI should of course be included as the PID.

2.1.7. Publication date

The publication date is the year of first publication of a dataset formatted as a four-digit number.

The year itself appears as a disambiguating element rather than an indication of the recency of the cited data. This important information may be signalled by the update date (see below), an access date or a 'most recent data' date.

Example: 2021

Where the dataset has been updated since first publication, the year should be followed by 'updated' and the publication date of last update in parentheses.

Example: 2013 (updated 2021-02-17)

For continuing datasets and datasets associated with recurrent surveys where the year associated with the data is contained in the dataset title, the date of first publication offers little information and may be omitted.

2.1.8. Date of citation

The date of citation expresses the date on which the creator of the citation accessed the data.

The date of citation may include one or more of the following elements that assist with the precision of changing datasets:

- 'accessed' followed by a date;
- 'most recent data' followed by a date (see Section 1.3.1);
- 'snapshot available as' followed by an actionable (clickable) PID (see Section 1.3.2.1);
- 'snapshot available at' followed by an http URI (web address) (see Section 1.3.2.1).

Where more than one of these elements is present, they should be separated by commas.

Section 2.3 is essential reading for more detail on the formatting of dates.

2.1.9. Persistent identifier

The PID of the dataset is assigned by the creators or publishers and managed within a recognised scheme, such as DOI.

A PID should, where possible, appear in an actionable (clickable) URI form so that a browser will take the user to the dataset. The URI should preferably link to a 'landing page' for the dataset rather than the dataset itself. The user can then review the metadata and ensure that loading the data itself is appropriate.

The URI should normally be taken verbatim from the metadata of the original dataset. However, if the identification scheme recommends a particular form, the URI should be adapted to this.

Example: DataCite now recommends that the https form of a URI should be used for its DOIs, so a URI in dataset metadata that reads:
http://doi.org/10.5281/zenodo.5501835
should appear in a citation as:
https://doi.org/10.5281/zenodo.5501835

Because the PID URI acts as both an identifier and an indication of where the dataset is available, it is not necessary to include 'available at' before the PID.

Where the dataset uses a PID from a family that is not widely known or trusted, or appears unlikely to be maintained, it may be appropriate also to include in the citation a direct link to the dataset as a backup.

When presented in formats that allow the presentation of hypertext links (such as HTML, Office Open XML document or PDF) the actionable URI should appear as both the anchor text and the destination of the link. The normal formatting (usually underlined blue text) should be retained.

Diligent checks should be made to find the appropriate identifier; however, if none is available, great care should be taken to ensure that the rest of the citation contains enough information to unambiguously define the cited data.

2.2. The citation itself

2.2.1. Editorial styles

To be useful, a citation must appear in the document (which is intentionally a rather general term) rather than somewhere else. A reader can then detect its presence and make use of its content. And to avoid ambiguity, there must be something in the text at the point where the citation is invoked. This 'anchors' the dataset that is cited to the context in which it is being called out.

There are different ways in which a citation can be documented. These are used in different circumstances according to the type of document in which it appears, the intended audience and other factors such as house style.

As well as formal citations, there are informal citations where, for example, a journalist uses data to write a story or create a chart for consumption by non-specialist readers. These are discussed in Section 2.2.2.1, but it is important to be realistic here. Mass-market publications have their own style guides, and externally imposed formats are unlikely to be respected. That said, the principles in this guide are important because, as several publishers and journalists have said, news is the first rough draft of history ([27]).

A formal citation can be placed:

- inline in the text at the point where the citation is made;
- as part of the caption of a figure, table or infographic;
- in a footnote, which normally appears at the foot of a page on which the citation is made, with some kind of anchor in the text that links to the footnote;
- in a reference list ([28]) that appears at the end of a document (or at the end of the chapter or section if the document is structured that way) – again, an anchor is needed to link to the entry in the reference list.

The choice between these four options is a matter of house style, or possibly of personal preference on the part of the author. Opinions of which is 'best' are entrenched and firmly held. Authors should not lose sight of the objective, which is to communicate the citation with precision while ensuring that the text remains readable and usable.

Where the work is to be published in a journal or examined for an academic qualification, the author will probably be required by house rules to use a particular style. Where the work is subject to the *Interinstitutional Style Guide* (ISG), the provisions there on references and footnotes are to be followed, and these are noted where appropriate.

Recommendation: Where a particular style of citation is mandated by a publisher or institution that style should be followed, but the best practice set out in this guide should be followed as principles insofar as they are compatible with the required style.

Recommendation: Where a citation style is to be chosen, authors should review the alternatives and choose the version that allows a reader to make the most productive use of the text while providing effective and unambiguous linking to the citations when they are to be accessed.

([27]) Often attributed to Washington Post publisher Phil Graham; earlier use by journalist Alan Barth is documented.

([28]) Note that editors make a distinction between a 'reference list', which contains the entities cited in the work, and a 'bibliography', which is a list of further reading recommended by the authors and possibly examined but not directly cited. These terms are sometimes used interchangeably, however.

User story: A paper contains a couple of simple citations of straightforward datasets. An inline style should be used because the citations do not disrupt the flow of most of the text.

User story: A report contains one or two data citations per page of typeset text. The use of a footnote style is appropriate because it occupies a small part of each page and the reader's eye can move between the text and the citation with ease.

User story: A policy submission contains hundreds of citations, with some pages carrying a high proportion of them. A reference list allows the citations to be grouped together away from the text so they do not distract but are still available and are linked to the text that references them.

2.2.2. Citation formats and examples

The elements defined in Section 2.1 allow the construction of citations in each of the available styles.

2.2.2.1. Journalistic

Ad hoc citation of data in less-formal publications will not in general be able to follow the strict guidelines in the rest of this section, but this does not mean that writers and editors should ignore the thinking behind them.

User story: A journalist writes a story about forestry production and uses data from Eurostat. As well as acknowledging the origin of the data, they mention the Eurostat dataset used, 'tag00073', so that readers can check it if they want to.

Recommendation: Where formal citation is not possible or appropriate, the principles in this guide should be taken into consideration as far as possible.

2.2.2.2. Inline

Because of the need to avoid disrupting the text, the information in an inline citation reflects a minimalist approach and, although it must be formally unique, it may usefully be augmented by textual information outside the citation itself. Its uniqueness comes from the use of a PID, and it should not normally be considered without one. The use of an inline reference should probably be considered a last resort, and the additional information in a footnote or reference-list entry may make those a better choice in many cases.

Recommendation: A citation that is placed inline in the document text should be constructed as: (authors, date, PID)

Example:
… access to bibliographical metadata of the EU general publications (Publications Office of the European Union, 2014, http://data.europa.eu/doi/10.2906/112117098108/4) is offered …

Example:
… analysis of the JRC-EU-TIMES model (Nijs, W. and Ruiz, P., 2019, http://data.europa.eu/89h/8141a398-41a8-42fa-81a4-5b825a51761b) shows that …

Example:
… and the national GDP measurements (European Commission, Eurostat, https://ec.europa.eu/eurostat/databrowser/bookmark/972688bc-2552-4201-b8fe-c9e514a352b4?lang=en) were plotted against …

Note the omission of the publication date.

2.2.2.3. Caption

When a document includes elements such as figures, graphs, diagrams or infographics, the source of the data used to create them should be closely associated with the element so that if the element is separated from the full document, the source is still clear.

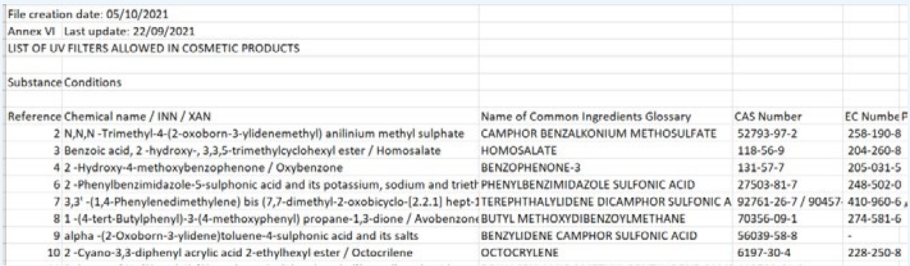
However, the issue noted in Section 1.2.1 of the citation being remote from explanatory text should be addressed here: where more than one data source is used, the nature of the processing to create, for example, an infographic should be made clear.

It is customary for the caption of the element to include information about the 'source' of the used data. That forms the basis of this recommendation.

Recommendation: A citation placed in the caption of a graphical element of a document should, if possible, be constructed in the same way as an inline citation, preceded by '***Source***':

Source: author(s), date, PID

Example:



File creation date: 05/10/2021

Annex VI Last update: 22/09/2021

LIST OF UV FILTERS ALLOWED IN COSMETIC PRODUCTS

Substance Conditions

Reference	Chemical name / INN / XAN	Name of Common Ingredients Glossary	CAS Number	EC Numbe
2	N,N,N -Trimethyl-4-(2-oxoborn-3-ylidenemethyl) anilinium methyl sulphate	CAMPHOR BENZALKONIUM METHOSULFATE	52793-97-2	258-190-8
3	Benzoic acid, 2 -hydroxy-, 3,3,5-trimethylcyclohexyl ester / Homosalate	HOMOSALATE	118-56-9	204-260-8
4	2 -Hydroxy-4-methoxybenzophenone / Oxybenzone	BENZOPHENONE-3	131-57-7	205-031-5
6	2 -Phenylbenzimidazole-5-sulphonic acid and its potassium, sodium and trietl	PHENYLBENZIMIDAZOLE SULFONIC ACID	27503-81-7	248-502-0
7	3,3' -(1,4-Phenylenedimethylene) bis (7,7-dimethyl-2-oxobicyclo-[2.2.1] hept-1	TEREPHTHALYLIDENE DICAMPHOR SULFONIC A	92761-26-7 / 90457-	410-960-6
8	1 -(4-tert-Butylphenyl)-3-(4-methoxyphenyl) propane-1,3-dione / Avobenzone	BUTYL METHOXYDIBENZOYLMETHANE	70356-09-1	274-581-6
9	alpha -(2-Oxoborn-3-ylidene)toluene-4-sulphonic acid and its salts	BENZYLIDENE CAMPHOR SULFONIC ACID	56039-58-8	-
10	2 -Cyano-3,3-diphenyl acrylic acid 2-ethylhexyl ester / Octocrilene	OCTOCRYLENE	6197-30-4	228-250-8

Source: European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 2016, http://data.europa.eu/88u/dataset/cosmetic-ingredient-database-2-list-of-substances-prohibited-in-cosmetic-products

Sometimes this will not be possible, and an alternative approach is needed.

Recommendation: When it is not possible to identify the source data with precision in this way, the caption should include an anchor to a footnote or reference-list item, as described below.

2.2.2.4. Footnote or reference-list entry

The format used for a citation placed in a footnote or reference list offers more detail than an inline citation and may sometimes be the only way of specifying the used data with sufficient precision. The two formats are identical, and differ only in their location in the document.

Previous systems presented more detail in the reference list, but the distinction between the two is becoming irrelevant as publishing systems enable the direct linking ('hyperlinking') of the text and footnote / reference-list entries.

Footnotes and reference-list entries should be numbered sequentially. When subject to the *Interinstitutional Style Guide* (ISG), the number appears at the beginning of the citation as a superscript between parentheses.

Recommendation: A citation that is placed in a footnote or reference-list entry should be constructed as:

$(^{\text{number}})$ author(s), 'title', version, publisher, date, date of citation, PID

Example:
$(^{8})$ European Commission, Directorate-General for Energy, 'Energy statistical datasheets for the EU countries', 2016, accessed 2020-10-29, http://data.europa.eu/88u/dataset/information-on-energy-markets-in-eu-countries-with-national-energy-profiles

Example:
([17]) Universität Karlsruhe – Institut für Werkstoffkunde I, Beck, T. and Rau, K., 'Thermo-mechanical fatigue test data for NIMONIC 90 sa material for a temperature range of 400 to 850 °C and a mechanical strain range of.7 %', version 1.0, European Commission, Joint Research Centre, 2017, https://data.europa.eu/doi/10.5290/610001

Example:
([6] Dottori, F., Alfieri, L., Salamon, P. et al., 'Flood hazard map of the World – 10-year return period', European Commission, Joint Research Centre, 2016, http://data.europa.eu/89h/jrc-floods-floodmapgl_rp10y-tif

Example:
([21]) European Commission, Eurostat, 'GDP and main components (output, expenditure and income)' (namq_10_gdp), accessed 2021-06-28T23:02, https://ec.europa.eu/eurostat/databrowser/bookmark/972688bc-2552-4201-b8fe-c9e514a352b4?lang=en

Note the omission of the publication date and the inclusion of a short code (namq_10_gdp) used by Eurostat to identify the dataset.

2.2.2.5. Citation anchor – *Interinstitutional Style Guide* numbered note style

An anchor must be placed in the text to create a link to the footnote or reference-list element. When subject to the *Interinstitutional Style Guide* (ISG), the numbering style of that guide must be followed.

Recommendation: When subject to the *Interinstitutional Style Guide*, the anchor referencing a footnote or reference-list entry appears as a superscript between parentheses, separated from the preceding text by a thin space.

Example:
… analysis of fatigue test data ([17]) showed that …

Example:
… and the national GDP measurements ([23]) were plotted against …

2.2.2.5.1. Citation anchor – other numbered note style

When not subject to the *Interinstitutional Style Guide* (ISG), other formats similar to that described in Section 2.2.2.5 may be used. These will typically be defined in other house style documents and may require different treatment of the footnote or reference-list entry number.

Example: Some styles require the footnote or reference-list entry number to be surrounded by square brackets.

2.2.2.6. Citation anchor – author–date

When using a reference list, an alternative approach to creating an anchor may be specified. Rather than numbering the citations, they are referenced in the text by the author and date. The corresponding entry in the reference list contains the same name and date, and the rest of the citation can be discovered.

Several widely used systems including the ***Chicago Manual of Style*** and the American Psychological Association's ***APA Style*** allow this scheme.

Where the year of publication is not useful and is omitted (see Section 2.1.7), it is replaced by a dataset code that appears in the full reference-list entry.

Recommendation: When the author-date system is adopted, the in-text anchor for a dataset citation should be constructed as:

(author(s), date) or (author(s), dataset code)

The following in-text anchors relate to the footnote and reference-list examples above.

Example:
… having normalised the available data (European Commission, Directorate-General for Energy, 2016), results were compared …

Example:
… and the national GDP measurements (European Commission, Eurostat, namq_10_gdp) were plotted against …

Note that in place of the year of publication, the short code appears to provide an effective link to the reference list.

Example:
… analysis of fatigue test data (Universität Karlsruhe – Institut für Werkstoffkunde I, Beck, T. and Rau, K., 2017) showed that …

2.3. Dates in citations

Citing books or journal articles by publication date does not require great precision. Because data can change, citing data is different, and the date and time of data instance, capture, publication and citation are critical.

The ISG contains rules on the formatting of dates and times in documents, but with respect to data portals in the European Union, DCAT-AP is an agreed specification for describing public-sector datasets. It allows interoperability between different data portals and ensures that data is reusable.

DCAT specifies a 'Data Catalog Vocabulary' [29] recommended by the World Wide Web Consortium (W3C). DCAT-AP [30] is a specification based on W3C's Data Catalog Vocabulary (DCAT) for describing public-sector datasets in Europe. Both the core specification and the application profile are under constant development, and users should ensure they are aware of the latest publications.

In DCAT, dates and times are to be constructed in accordance with a W3C note, 'Date and time formats' [31], which defines a profile of the ISO standard [32] for dates and times.

2.3.1. *Interinstitutional Style Guide* date and time format

The ISG formats may be used where backward compatibility is important, but for most data citation purposes the DCAT-AP indicated datatypes should be applied.

The ISG allows a full date format with the month spelled out.

Example: 4 September 2020

The ISG also allows an abbreviated format using the month number, but this is not recommended for data citations because of the ambiguity between 2 June 2021 and 6 February 2021, either of which might (in different cultures) be written as, or understood from, '6.2.2021'.

Example to avoid: 4.7.2020

[29] https://www.w3.org/TR/vocab-dcat-2/

[30] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/201-0

[31] https://www.w3.org/TR/NOTE-datetime

[32] The W3C note references ISO 8601:1988 'Data elements and interchange formats – Information interchange – Representation of dates and times'. This standard has long been superseded and the current version is ISO 8601-1:2019 'Date and time – Representations for information interchange – Part 1: Basic rules'. There are few important differences that affect this guide except that 24:00 is no longer a valid time (it should be written as 00:00 of the following day).

The ISG allows times to be specified using the 24-hour system, and requires hours and minutes to be separated by a point and a leading zero in the hour field to be omitted.

Example: 9.23

Example: 17.47

The ISG also allows the use of the 12-hour system with a.m. or p.m., but this is not recommended for data citation.

Example to avoid: 7.34 p.m.

The ISG does not specify how to indicate time zones, and where the exact time of data access is important the DCAT-AP formats should be used.

2.3.2. Date and format compliance with DCAT-AP (xsd:date or xsd:dateTime)

The use of the DCAT-AP date and time formats is generally recommended for data citation. DCAT-AP uses primitive datatypes xsd:date or xsd:dateTime, inspired by ISO 8601:1988 (33). The use of a standard format is important for interoperability, which DCAT-AP supports. Although ISO 8601-1 offers numerous options, with the intention that particular applications select an appropriate one, W3C and consequently DCAT and DCAT-AP offer a single system for simplicity and certainty.

DCAT-AP adopts a scalable approach to precision: elements representing a greater precision than is warranted can be omitted without loss of interoperability. Examples appear below.

Recommendation: In citations of data, dates and times should, whenever possible, use the DCAT-AP indicated datatypes and adopt a level of precision appropriate to the case.

2.3.3. Time zones

Specifying a time without reference to its geography is ambiguous because different absolute times are implied in different time zones. Using Coordinated Universal Time (UTC) – broadly the same as Greenwich Mean Time – avoids this.

Recommendation: Where a date and time is specified more precisely than a whole day, the applicable time zone should be stated, and should preferably use UTC and be designated by a time-zone designator of 'Z'. In exceptional circumstances a local time zone may be specified using the appropriate time-zone designator (see example).

[33] https://www.w3.org/TR/xmlschema-2/#isoformats

2.3.4. International standard date and time examples

The following examples show how ISO 8601-1 can specify time and date to different levels of precision.

Examples

2017	The whole year 2017
2017-02	The month of February in 2017
2017-02-12	12 February 2017
2017-02-12T14Z	2 p.m. UTC on 12 February 2017
2017-02-12T14:27Z	2.27 p.m. UTC on 12 February 2017
2017-02-12T14:27:15Z	2.27 and 15 seconds p.m. UTC on 12 February 2017
2017-02-12T14:27:15-05	2.27 and 15 seconds p.m. US Eastern Standard Time on 12 February 2017

2.3.5. Other formats

Other date formats should in general be avoided because of the possible ambiguity between June 2, 2021 and February 6, 2021, either of which might be written as, or understood from, 6/2/2021 or 6.2.2021. While '7 Dec 2020' is unambiguous, the benefits of a slightly shorter citation than '7 December 2020' are outweighed by the loss of consistency.

Recommendation: The use of date and time formats other than the selected ISG format and DCAT-AP-indicated datatypes should be avoided.

3. PART THREE – Other information

This final part summarises areas where the recommendations in this guide differ from other documents so that variances can be understood, and provides a select list for further reading.

1.1. Differences between recommendations

This guide presents recommendations for a relatively new activity; administrators and scholars have been citing text documents for centuries, but data for only a few years. There is no universal consensus in this field, and these recommendations diverge from other norms in a few ways.

Consequently, no unique citation style has yet been approved by the EU institutions: different styles have developed in different bodies as they respond to different needs. The requirements of scientists using data from the European Commission's Joint Research Centre differ from those of statisticians using data from Eurostat, for example.

As the field develops, this guide may provide a foundation for further harmonisation of styles.

1.2. Basis for recommendations

Two documents contain key recommendations that have informed the contents of this guide.

ISO 690:2021 'Information and documentation – Guidelines for bibliographic references and citations to information resources' has recently been revised to combine paper and electronic publications. It now provides guidance on citing non-traditional resources – including datasets, which it refers to as 'research data'. However, ISO 690 does not attempt to define a 'style' – its recommendations cover the content of a citation rather than its presentation. This guide attempts to follow ISO 690, though some exceptions are noted below.

The *Interinstitutional Style Guide* (ISG) is published by the OP. It aims to ensure consistency and precision in documents generated within the EU. It does not contain specific guidance on citing data, but it does offer a general structure for citations (which it terms 'references'). The recommendations of the ISG should be followed in documents created by or on behalf of the EU institutions.

This guide does not form part of the ISG but recommends a style that is consistent with it.

1.3. Specific differences

The following differences may be noted between the recommendations noted above and those in this guide.

3.3.1. Omission of publisher

ISO 690:2021 and previous EU institution practice requires the presence of both the corporate author and the publisher of a dataset. In many cases these are the same entity, identified in a citation by the same name. The inclusion of a second instance of the same name adds no information, and this guide allows the omission of the publisher in these circumstances.

3.3.2. Omission of publication date

Where a dataset is continuously updated, the original publication date is of little value, and this guide recommends its omission as potentially misleading. The dataset must be accurately defined by other information (principally by a PID) to make up for the absence of this information.

3.3.3. Inclusion of author identifiers

ISO 690:2021 requires the inclusion of PIDs for corporate and individual authors, and for publishers. This guide requires this only when ambiguity would otherwise arise. Practice in this area is evolving, and the recommendation may change in the future.

3.3.4. Most recent data

With a continuing dataset, there are several important dates: the date the dataset was accessed, the date the dataset had last been updated when accessed and, in the case of time-series data (which will typically be the case for continuing datasets), the date associated with the most recent addition. For many purposes, the last of these is the most important, but existing guides have no way to specify it.

For this reason, this guide allows a 'most recent data' term to be used in the same way as 'accessed', recording the date associated with the most recent data available at the time of access.

3.3.5. File size

ISO 690:2021 requires that the file size be quoted in the citation 'if large'. This guide does not recommend that approach because practitioners dealing with datasets are used to dealing with 'large' files containing data. Furthermore, where dynamic data is cited, its size may vary over time.

Further reading (bibliography)

International Organization for Standardization, ISO 690:2021 'Information and documentation – Guidelines for bibliographic references and citations to information resources', Geneva, 2021.

International Organization for Standardization, ISO 8601-1:2019 'Date and time – Representations for information interchange – Part 1: Basic rules', Geneva, 2019.

Fenner, M., Crosas, M. and Grethec, J., 'A data citation roadmap for scholarly data repositories', 2017, https://doi.org/10.1101/097196

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D. et al. 'A data citation roadmap for scientific publishers', 2018, https://doi.org/10.1038/sdata.2018.259

Publications Office of the European Union, *Interinstitutional Style Guide*, http://publications.europa.eu/code/en/en-000100.htm

Glossary

This glossary is not comprehensive, but it includes various terms used in this guide and the meanings they have herein.

Altmetrics. Measures of impact of a scholarly output that take into account more than the traditional number of citations in other outputs. This may include views online and social media traffic.

Author. Person or organisation responsible for creating a dataset. Sometimes called a 'contributor'.

Bibliography. List of information sources that an author considers potentially relevant to an output, whether or not they are cited in it. Compare with 'Reference list'.

Citation. Process of indicating with precision what sources have been used to create content.

Data. Abstract concept of quantified information available for use in analysis.

data.europa.eu. The official portal for European data providing a single point of access to open data from international, EU, national, regional, local and geo data portals (https://data.europa.eu/en).

Dataset. Specific data collected together and made available as a named group.

DCAT. Data Catalog Vocabulary is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the web (https://www.w3.org/TR/vocab-dcat-2/).

DCAT-AP. DCAT Application Profile for data portals in Europe is a specification based on DCAT for describing public-sector datasets in Europe. Its basic use case is to enable a cross-data portal search for datasets and make public-sector data better searchable across borders and sectors (https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe).

Eurostat. The statistical office of the European Union, publishing official, harmonised statistics.

Hosting platform. Online service that allows scientific and scholarly outputs (including datasets) to be uploaded and made available to others.

Hyperlink. Element in text that can be clicked (or otherwise selected) to give direct access to linked information.

Identifier. String of characters or similar that denotes something or can stand in its place.

ISG – *Interinstitutional Style Guide*. Document setting out rules for the editing and presentation of documents within the European Union institutions.

JRC – Joint Research Centre of the European Commission. Part of the European Commission that undertakes research in support of EU policies.

OP – Publications Office of the European Union. Interinstitutional body within the EU that acts as the official publisher and archivist of EU materials in various formats.

Persistence. Property of information that means it will still be available to users when needed in the future.

PID – persistent identifier. Identifier that is designed to be persistent, or least to be able to be persistent if well managed.

Plagiarism. Copying or otherwise including work generated by someone else without acknowledgement.

Precision. In citation, the specification of the source information in such a way that a reader can understand exactly and without ambiguity the information that was used by the writer.

Publisher. Organisation (or occasionally person) that makes a dataset available.

Reference list. List of information sources that have been used in generating a particular output. Compare with 'Bibliography'.

Reproducibility. Within the scientific method, the ability for the same result to be independently obtained by other researchers.

Reuse. Use for some purpose of outputs generated for a different purpose.

Zenodo. Widely used repository for research papers, data and other outputs operated by CERN as a 'marginal activity'.

Annex 1 – Checklist (long) for footnotes or reference-list entries

The following steps should be considered to be a checklist when creating a citation that will appear in a footnote or reference-list entry.

The key initial step is to refer to the metadata of the dataset that has been used. This may include a ‘cite this dataset as’ function. This will provide most or all of the information needed, though it may appear in a slightly different format and some editing work may be needed.

Otherwise, the metadata should include all the information needed, though sometimes a little ‘research’ will be needed to locate some of it.

author(s)	Who does the cited dataset note as its creator or as a contributor to it? Is there a corporate author as well as the individual authors? How many are there and in what order do they appear in the metadata of the original dataset? Decide how many to include if there are more than three – this guide recommends a maximum of three (followed by ‘et al.’) but allows the inclusion of fewer or more in some circumstances. Check whether the authors have been presented with their family name first (especially important with names from cultures that are not familiar to you).
‘title’ or *title* (short code)	What is the accurate title (including capitalisation and punctuation) of the original dataset? Is it necessary to capitalise the title to meet ISG requirements (when the title is presented in italics)? Is there a short code by which the dataset is known, which can be placed in parentheses after the title?
version	What is the version of the dataset that was used in the activity for which it is being cited? Do not just pick up the most recent version unless this is what was used and is appropriate in the citation.
publisher	Who made the dataset available? They might already be included as an author, in which case leave them out here. Don’t just assume that a repository is the publisher – there may be an actual publisher in the metadata.
date	When does the metadata say the data was published? For datasets that change, the date of initial publication may not be useful. However, sometimes it helps to distinguish similar but slightly different datasets. Is that the case here? Write the date (and time if needed) in datatypes indicted by DCAT-AP ([34]) (e.g. 2021-09-13 or 2021-09-13T23:16) or in the *Interinstitutional Style Guide* ([35]) format as appropriate.
date of citation ‘accessed’	If the data may have changed since it was accessed, or may change in the future, at least note the date it was accessed.
PID	There should be a persistent identifier in the dataset metadata. This will be managed so that people can still use it to find and access the data in the future. Try to locate it and, if possible, write it as a web link so that it can be clicked to access a landing page. Check that the web link works and gives the expected result before including it in a citation – a link that is already broken is of little use to readers.

[34] ‘DCAT Application Profile for data portals in Europe’, https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

[35] Publications Office of the European Union, *Interinstitutional Style Guide*, http://publications.europa.eu/code/en/en-000100.htm

Annex 2 – Checklist (short) for footnotes or reference-list entries

□ **author(s)**

□ **title (short code)**

□ **version**

□ **publisher**

□ **date**

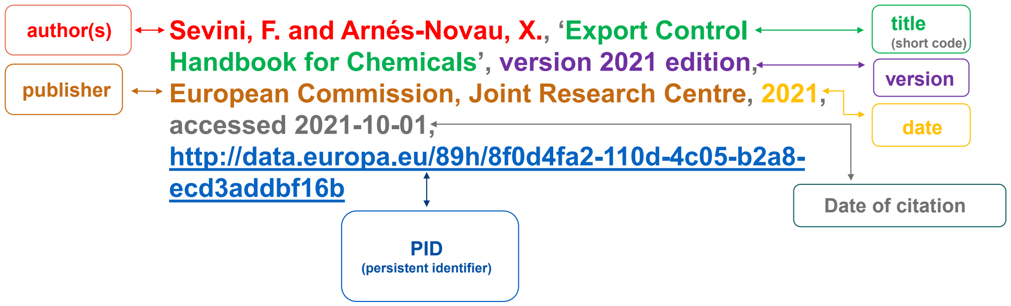
□ **date of citation**

□ **PID (persistent identifier)**

Elements:

Author(s), Title (short code), Version, Publisher, Date, Date of citation, PID (persistent identifier)

Annex 3 – Diagram for footnote and reference citation style with date format compliant with DCAT-AP

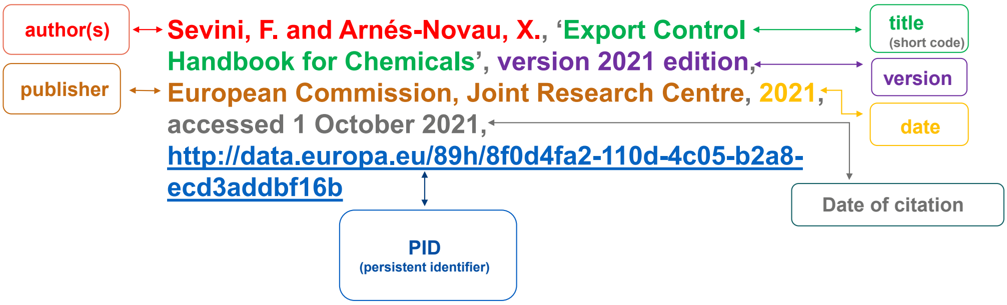


Sevini, F. and Arnés-Novau, X., 'Export Control Handbook for Chemicals', version 2021 edition, European Commission, Joint Research Centre, 2021, accessed 2021-10-01, http://data.europa.eu/89h/8f0d4fa2-110d-4c05-b2a8-ecd3addbf16b

Elements:

Author(s), Title (short code), **Version, Publisher, Date,** Date of citation, **PID (persistent identifier)**

Annex 4 – Diagram for footnote and reference citation style with date format compliant with the *Interinstitutional Style Guide*



Sevini, F. and Arnés-Novau, X., 'Export Control Handbook for Chemicals', version 2021 edition, European Commission, Joint Research Centre, 2021, accessed 1 October 2021, http://data.europa.eu/89h/8f0d4fa2-110d-4c05-b2a8-ecd3addbf16b

Elements:

Author(s), Title (short code), Version, Publisher, Date, Date of citation, PID (persistent identifier)

Annex 5 – Examples of citations with date format compliant with DCAT-AP

European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union, 'Consolidated list of persons, groups and entities subject to EU financial sanctions', 2016 (updated 2020-10-23), accessed 2021-10-01, http://data.europa.eu/88u/dataset/consolidated-list-of-persons-groups-and-entities-subject-to-eu-financial-sanctions-fisma

European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 'Tenders Electronic Daily (TED) (csv subset) – public procurement notices', version 3.3, 2015 (updated 2020-11-22), http://data.europa.eu/88u/dataset/ted-csv

European Commission, Directorate-General for Informatics, 'National Interoperability Framework Observatory (NIFO) – Digital Public Administration factsheets 2020', accessed 2021-10-01, http://data.europa.eu/doi/10.2906/100105103105116/1

European Commission, European Chemicals Agency, 'Mammalian toxicokinetic database (MamTKDB) 1.0', version 1.0, 2021, accessed 2021-10-01, http://data.europa.eu/88u/dataset/mammalian-toxicokinetic-database-mamtkdb-1-0

European Commission, European Centre for Disease Prevention and Control, 'COVID-19 coronavirus data – daily (up to 14 December 2020)', accessed 2021-10-01, http://data.europa.eu/88u/dataset/covid-19-coronavirus-data-daily-up-to-14-december-2020

European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 'Cosmetic ingredient database (Cosing) – List of substances prohibited in cosmetic products', 2016 (updated 2018-12-14), http://data.europa.eu/88u/dataset/cosmetic-ingredient-database-2-list-of-substances-prohibited-in-cosmetic-products

'European database of suspected adverse drug reaction reports (EudraVigilance)', European Commission, European Medicines Agency, 2015 (updated 2019-01-10), http://data.europa.eu/88u/dataset/suspected-adverse-drug-reaction-reports

European Commission, Joint Research Centre, 'European Soil Database v2.0 (vector and attribute data)', 2021, accessed 2021-10-01, http://data.europa.eu/89h/jrc-esdac-1

Sevini, F. and Arnés-Novau, X., 'Export Control Handbook for Chemicals', version 2021 edition, European Commission, Joint Research Centre (JRC), 2021, accessed 2021-10-01, http://data.europa.eu/89h/8f0d4fa2-110d-4c05-b2a8-ecd3addbf16b

Nijs, W. and Ruiz, P., '01_JRC-EU-TIMES Full model', European Commission, Joint Research Centre, 2019, http://data.europa.eu/89h/8141a398-41a8-42fa-81a4-5b825a51761b

'AMECO – ECFIN annual macroeconomic database', European Commission, Directorate-General for Economic and Financial Affairs, (updated 2020-11-05), http://data.europa.eu/88u/dataset/ameco

European Commission, Directorate-General for Energy, 'Energy modelling – EU Reference Scenario 2016', http://data.europa.eu/88u/dataset/energy-modelling

European Commission, Eurostat, 'Airport traffic data by reporting airport and airlines' (avia_tf_apal), most recent data 2021-09-01, https://ec.europa.eu/eurostat/databrowser/view/avia_tf_apal/default/table?lang=en

European Commission, Eurostat, 'Total length of motorways' (ttr00002), accessed 2021-10-15, https://ec.europa.eu/eurostat/databrowser/view/ttr00002/default/table?lang=en

European Commission, Eurostat, 'Real GDP growth rate – volume' (tec00115), updated 2021-09-28, https://ec.europa.eu/eurostat/databrowser/view/tec00115/default/table?lang=en

Publications Office of the European Union, 'Country Named Authority List', 2009 (updated 2021-09-29), http://data.europa.eu/88u/dataset/country

Annex 6 – Examples of citations with date format compliant with the *Interinstitutional Style Guide*

European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union, 'Consolidated list of persons, groups and entities subject to EU financial sanctions', 2016 (updated 23 October 2020), accessed 1 October 2021, http://data.europa.eu/88u/dataset/consolidated-list-of-persons-groups-and-entities-subject-to-eu-financial-sanctions-fisma

European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 'Tenders Electronic Daily (TED) (csv subset) – public procurement notices', version 3.3, 2015 (updated 22 November 2020), http://data.europa.eu/88u/dataset/ted-csv

European Commission, Directorate-General for Informatics, 'National Interoperability Framework Observatory (NIFO) – Digital public administration factsheets 2020', accessed 1 October 2021, http://data.europa.eu/doi/10.2906/100105103105116/1

European Commission, European Chemicals Agency, 'Mammalian toxicokinetic database (MamTKDB) 1.0', version 1.0, 2021, accessed 1 October 2021, http://data.europa.eu/88u/dataset/mammalian-toxicokinetic-database-mamtkdb-1-0

European Commission, European Centre for Disease Prevention and Control, 'COVID-19 coronavirus data – daily (up to 14 December 2020)', accessed 1 October 2021, http://data.europa.eu/88u/dataset/covid-19-coronavirus-data-daily-up-to-14-december-2020

'European database of suspected adverse drug reaction reports (EudraVigilance)', European Commission, European Medicines Agency, 2015 (updated 10 January 2019), http://data.europa.eu/88u/dataset/suspected-adverse-drug-reaction-reports

European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 'Cosmetic ingredient database (Cosing) – List of substances prohibited in cosmetic products', 2016 (updated 14 December 2018), http://data.europa.eu/88u/dataset/cosmetic-ingredient-database-2-list-of-substances-prohibited-in-cosmetic-products

European Commission, Joint Research Centre, 'European soil database v2.0 (vector and attribute data)', 2021, accessed 1 October 2021, http://data.europa.eu/89h/jrc-esdac-1

Sevini, F. and Arnés-Novau, X., 'Export control handbook for chemicals', version 2021 edition, European Commission, Joint Research Centre (JRC), 2021, accessed 1 October 2021, http://data.europa.eu/89h/8f0d4fa2-110d-4c05-b2a8-ecd3addbf16b

Nijs, W. and Ruiz, P., '01_JRC-EU-TIMES full model', European Commission, Joint Research Centre, 2019, http://data.europa.eu/89h/8141a398-41a8-42fa-81a4-5b825a51761b

'AMECO – ECFIN annual macroeconomic database', European Commission, Directorate-General for Economic and Financial Affairs, (updated 5 November 2020), http://data.europa.eu/88u/dataset/ameco

European Commission, Directorate-General for Energy, 'Energy modelling – EU reference scenario 2016', http://data.europa.eu/88u/dataset/energy-modelling

European Commission, Eurostat, 'Airport traffic data by reporting airport and airlines' (avia_tf_apal), most recent data 1 September 2021, https://ec.europa.eu/eurostat/databrowser/view/avia_tf_apal/default/table?lang=en

European Commission, Eurostat, 'Total length of motorways' (ttr00002), accessed 15 October 2021, https://ec.europa.eu/eurostat/databrowser/view/ttr00002/default/table?lang=en

European Commission, Eurostat, 'Real GDP growth rate – volume' (tec00115), updated 18 September 2021, https://ec.europa.eu/eurostat/databrowser/view/tec00115/default/table?lang=en

Publications Office of the European Union, 'Country named authority list', 2009 (updated 29 September 2021), http://data.europa.eu/88u/dataset/country

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696, or
— by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from: https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The official portal for European data (https://data.europa.eu/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Publications Office
of the European Union

ISBN 978-92-78-42673-6