Data citation practices across earth, life, social sciences and humanities - opportunities for OpenAIRE.

Contributors:
Ben Companjen, **Maarten Hoogerwerf**, Data Archiving and Networked Services, DANS, NL. **Paolo Manghi**, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR, Pisa, Italy. **Florian Graef**, **Jo McEntyre**, European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. **David Bousfield**, Cambridge, UK. **Uwe Schindler**, **Michael Diepenbroek**, PANGAEA, MARUM, University of Bremen, Germany.

Table of Contents

Motivation for making data citation work better

Citing data sources used in research is one of the key requirements for effective scientific progress. In the era of Open Science (Boulton et al., 2012), citing data becomes a key integration mechanism between research articles and the underlying data resources, whether they are new data generated in the course of the study being described, or existing data in public databases that have been recited or reused in some way. Citing data provides provenance for scientific assertions as well as credit for the generation of the dataset(s). As new methods for generating large quantities of data emerge, combined with open access policies and mandates for articles and data, the vision of an open information infrastructure across disciplines becomes a viable reality.

However to make transactions in this network navigable for both humans and machines, many of the practices around citing data will need to be improved. Over the next few years, it will become unacceptable to just mention data within narrative sentences, or provide low-resolution images of results without either the underlying data or the high resolution files. Several initiatives involving both community outreach and technical implementation are

building momentum to support rigour in data citation (see Appendix), however we are still at the beginning of this process, and many challenges remain both within and across different scientific disciplines, which may also have different existing practices for data citation. For many stakeholders (for example, data creators, data publishers/centres, researchers, funders, journal/book publishers and institutions) the goal is to make data citation a globally integrated enterprise. The purpose of this document therefore is to provide an overview of the rapidly evolving and diverse landscape of data citation and define the role of OpenAIRE within it.

Data citation and the OpenAIRE context

The OpenAIRE2020 proposal for this Task was to hold a workshop on the area of data citation. However, when discussing what to focus on, we realised that while there are many ongoing efforts in this area, there was not a great deal of shared understanding or cohesion on what these might mean in different subject contexts, given that the current status of data citation in different disciplines is not clear even within and across the small subset represented by the partners in this Task.

We recognise that the partners in this OpenAIRE task come from various backgrounds. Each partner wishes to improve data citation practices in their own disciplinary context. Therefore the first step is to share use case information on current status and developments around data citation, ensuring joint understanding. In this way, we hope to identify key actions that can be discussed more widely within the OpenAIRE community. Secondly, as a group, the partners on this task should share information on the wider efforts of facilitating data citation to inform and avoid duplication of effort:

- In the life sciences, the partners want to see data cited in both human and machine-actionable ways during the publication process. The goal is therefore to develop tools that support structured data citation using the Journal Article Tag Suite (JATS) in publishing workflows. Data is mentioned in articles relatively commonly, and some effort will be required to distinguish de novo data citation from reuse, which in turn can provide insight into the impact of particular datasets or collections.
- In the social sciences and humanities in the Netherlands, Data Archiving and Networked Services (DANS) wants to make sure that they - as an archive and national portal - can capture citation-information properly when datasets are ingested into their archive, or when metadata is harvested from institutional repositories into their portal, preserve these and publish these for usage by consumers like OpenAIRE.
- In the earth sciences, authors mainly refer to literature sources in their work and rarely cite data. Therefore, data centers need to keep track of the links from the datasets back to the articles. Furthermore, the data-literature linking service is important for literature publishers so that they can annotate article metadata using the “reverse links” to data. The goal is to make it easier for scientists to directly cite datasets and give credit to the data producers. Data publication needs to become an integral part of the scholarly publishing workflow.

Thus we need to ensure that our initiatives complement each other, so that we can reach out to the various stakeholders in the OpenAIRE community consistently.

With respect to adoption and implementation, we expect this process to highlight opportunities to use similar approaches or standards across disciplines. This aspiration is aligned well with the OpenAIRE mission and we expect to identify important actions as a result. Suggestions for future work are outlined in the final section of this document: “Outcomes and next steps”.

Status quo of citing data: disciplinary perspectives

This section outlines the status quo of data citation for Earth Sciences, for Life Sciences and for Social Sciences and Humanities.

Earth Sciences [PANGAEA]

The Publishing Network for Geoscientific & Environmental Data (PANGAEA) is an Open Access library archiving, publishing, and distributing geo-referenced data about climate variability, the marine environment and geological research. It provides an archive for any kind of data from earth system research and thus has no special format requirements for submissions. However, for samples taken or measurements made somewhere on earth, the provision of position(s) is mandatory (latitude/longitude in decimal preferred).

The system is hosted by the Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research (Bremerhaven, Germany) and the Center for Marine Environmental Sciences (MARUM, University of Bremen, Bremen, Germany). It is supported with funding from The European Commission, Research, Federal Ministry of Education and Research (BMBF), Deutsche Forschungsgemeinschaft (DFG), International Ocean Discovery Program (IODP)

PANGAEA is a member of the ICSU World Data System. Its policy for data management and archiving follows the ICSU World Data System Data Policy and the OECD Principles and Guidelines for Access to Research Data from Public Funding. Authors submitting data to the Pangaea data library for archiving must agree that all data are provided under a creative commons license.

Most of the data are freely available and can be used under the terms of the license mentioned on the data set description. A few password protected data sets are under moratorium from ongoing projects. The description of each data set is always visible and includes the principle investigator (PI) who may act as a gatekeeper for access. Each dataset can be identified, shared, published and cited by using a Digital Object Identifier (DOI). Data are archived as supplements to publications or as citable data collections. Citations are available through the portal of the German National Library of Science and Technology. Archiving follows the Recommendations of the Commission on Professional Self Regulation in Science for safeguarding good scientific practice.

The PANGAEA data model reflects the standard hierarchy of activities involved in geoscience data collection. Any type of information, data and documents may be served (profiles, maps, photos, graphics, text and numbers). Thus a Project is composed of expeditions for sampling. During an expedition samples may be taken or measurements made at a number of physical locations. The substances sampled such as rock, sediment, water, ice, can be sub-sampled or analysed. Finally these data are organised into datasets annotated by further information about methods used, parameters involved and so forth. Data are stored in a relational database in a consistent format with related meta-information following international standards. Data are geo-referenced in space and/or time so that individually configured subsets may be extracted.

PANGAEA is a designated archive for the journal Earth System Science Data (ESSD) published by Copernicus Publications and various other journals related to earth system research. The search engine is powered by the open-source software Elasticsearch and metadata processing is provided by panFMP *(PANGAEA Framework for Metadata Portals)*. Data sets deposited in PANGAEA will be automatically linked to corresponding articles in Elsevier's ScienceDirect. Articles linked in this way can also be linked to a Google Map displaying the geographical locations from which related PANGAEA data sets can be found.

By using PANGAEA as a publisher, data entities, as defined by the authors, are handled in a similar manner to scientific publications. Each dataset's metadata record starts with a citation consisting of standard bibliographic fields:

- Author(s); last name and first name/initial, separated by comma
- Year of publication, in brackets, i.e. the year when the data set was made available (digital or print);
- Title of the data set;
- Source institution; this information is optional and only used, if data are not supplementary to a publication;
- DOI, starting with `10.1594/PANGAEA.`

After technical review by the curator, import and approval of the author/PI, a dataset is set to "status published" and appears as citable data publication on the web page. In addition to submission of metadata and minting of the DOI at DataCite, this also initiates an entry in the library catalogue of the Technical Information Library in Hannover and thus becomes incorporated into the global library catalogues. For 4 weeks, this entry can still be edited until final DOI-minting. An immediate DOI minting can be forced by the curator on request. Similar to text publications, data publications can not be changed after final minting/publication.

PANGAEA supports three versions of citable data sets:

1) Data supplement - data are supplementary to a scientific paper and thus are an integral part of the paper and of its peer-review, e.g.:

> *Hollis, Chris (1993): Latest Cretaceous to Late Paleocene radiolarian biostratigraphy: A new zonation from the New Zealand region, Marine Micropaleontology, 21(4),*

295-327, doi:10.1016/0377-8398(93)90024-R , [supplementary data: doi:10.1594/PANGAEA.683866].

Following the rules for good scientific practice data need to be cited when being used. The full citation given by PANGAEA contains two DOIs: one linking to the paper itself, and the second as the DOI of the data supplement. There is an ongoing discussion depending on the scientific discipline whether a data supplement should be seen as a publication on its own (librarian view) or as part of the paper (scientific view).

Another point of discussion is the publication year of a data supplement. In new publications both years, the one for the paper and the one for the data supplement, are the same. When archiving old/printed data, the year of the data supplement can be the same as the year of the related publication or the year when the data were made available in digitized form (librarian view). Preference is given to the year of the printed version (to avoid confusion), e.g. doi:10.1594/PANGAEA.849173 .

2) Data publication that is not linked directly to the publication of an article can be supported through the publishing functionality of PANGAEA. Publishing data just through a data system is new and was invented during the DFG-project STD-DOI. This variant is not an established publication type in science so far. In pure data publications details of the publishing institute is included in the source information, e.g.:

Bauer, Wilfried; Spaeth, G; Jacobs, Joachim; Weber, Klaus; Siegesmund, S; Thomas, RJ (2004): Geological map of BJORNNUTANE, Heimefrontfjella, Antarctica (Scale 1:25 000). Alfred Wegener Institute for Polar and Marine Research, Bremerhaven; Institut for Geologie und Dynamik der Lithosphäre, Göttingen; Bundesamt für Kartographie und Geodäsie, Frankfurt, doi:10.1594/PANGAEA.138777

3) For peer-reviewed data publications various Data Journals for different scientific fields are available. Examples:

- Earth System Science Data by Copernicus,
- Scientific Data by Nature,
- Geoscience Data Journal by Wiley.

It is highly recommended that the data are referred to from articles with a full citation, see the list of data references on the following article:

Pesant, S. et al. (2015): Open science resources for the discovery and analysis of Tara Oceans data. Sci. Data 2:150023, doi:10.1038/sdata.2015.23

As data citations from literature as exemplified above are rare, PANGAEA is responsible for keeping track of the link from the data sets back to the articles (“reverse links”). In this context the data-literature linking service is important.

Life Sciences [European Bioinformatics Institute]

The Protein Data Bank (PDB) and the EMBL Data Library (later the European Nucleotide Archive, ENA) were the first internationally supported data repositories in the life sciences. In the case of ENA, the goal was to establish a publicly accessible central database of curated DNA and RNA sequences, rather than have scientists literally type out the sequences in journal articles. Since the 1990s, several deposition have been created to support many heterogenous data types. For example, the 2015 *Nucleic Acids Research* Database Issue contains 172 papers that include descriptions of 56 new molecular biology databases, and updates on 115 databases whose descriptions have been previously published in *NAR* or other journals. Many of these repositories are regulated at an international level (e.g. INSDC) and together with the primary research literature form a richly-networked global bioinformatics infrastructure. The European Bioinformatics Institute (EMBL-EBI) hosts many of such resources, which are frequently built in collaboration with other organizations internationally and integrated in scientifically useful ways.

The standards on which these data resources rely are created by community consensus and with the input of agencies such as the EMBL-EBI and NCBI in the USA. Several controlled vocabularies and ontologies are also in widespread use (e.g. MeSH, GO) that provide further consistent integration across resources.

Typically, when an author in the life sciences wishes to publish his or her work, the journal will require the author to deposit certain types of data (such as DNA sequences) in the appropriate public data resource. The value of depositing data of a specific type in a specific resource is that collectively, those data can be validated, curated, Evidence that this has indeed been done is usually the citation of a database accession number(s) (or PID(s)) within the body of the article. Such data archiving policies are policed with varying degrees of effectiveness by the scientific community and by the publishers themselves. The publication of an accession id is proof of compliance, and the deposition of the data in a single consensus archive avoids unnecessary replication, enables data curation and validation, and promotes data re-use (which for some data types, e.g. nucleotides, occurs on a huge automated scale). It is very common for submission data such as nucleotides to be reused in added-value resources and other interfaces that help users and save them time in finding useful information.

In articles, data are “cited” in a number of ways, including:

- unstructured mentions of identifiers or resources in the main text, or supplemental material
- structured reference in text, supplemental material
- data cited in reference lists

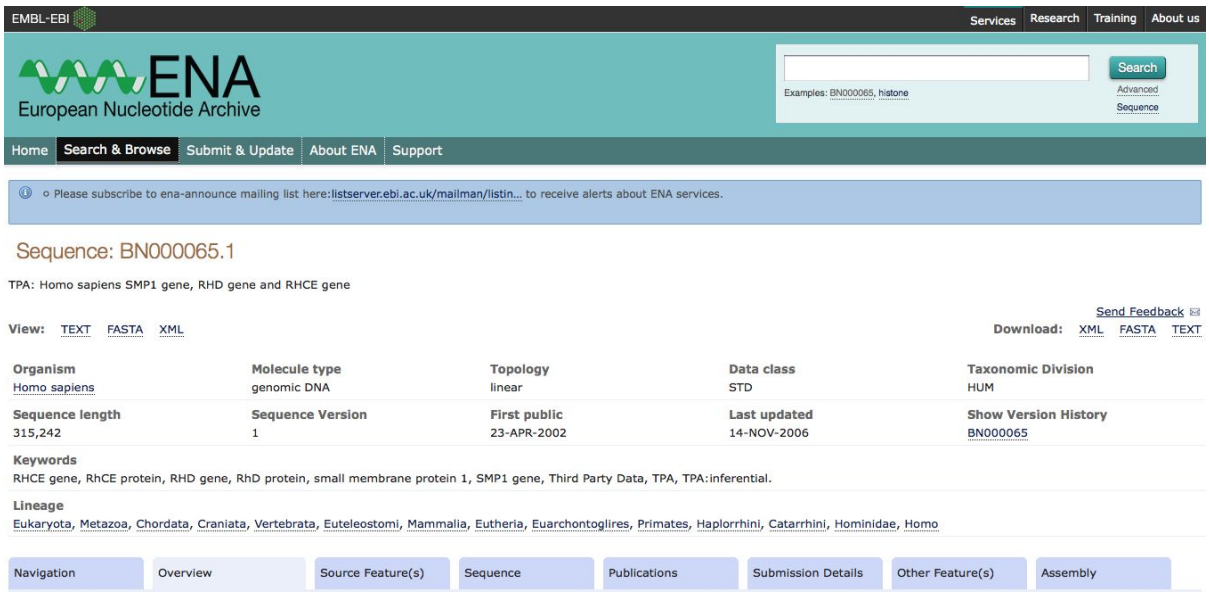
Ideally these links need to be formalised at the time of publication. To this end the Journal Article Tag Suite (JATS) is a standard that defines a set of XML elements and attributes for tagging various types of journal article (NLM DTD) and the data elements cited therein.

JATS is heavily used in life sciences, but needs to be extended in order to provide fully comprehensive facilities for data tagging.

An example: ENA accepts sequence data submissions intended for public release. During the process submitters have the opportunity to decide whether the data should become immediately public or should remain confidential for up to two years. It is most common for data to be released immediately, or on publication of the accompanying article. Once data has been publicly released it can be withdrawn from public access only in exceptional circumstances.

Typical metadata associated with such a submission will include: Accession_id, Accession_entity_name, Primary_reference_id (PMID), PMID_publication_date, Deposition_date, First_public_date, Version_history. This information is usually presented as a flat XML file whose structure varies significantly across and within the different repositories. Further subject-specific metadata provides deep annotation for further human and computational analysis.

Below we show an example of a database record for data accession BN000065 in the European Nucleotide Archive. Tabs at the bottom of the record lead to related information in various categories.



This “repository” scenario applies best to the “omic” disciplines, e.g. proteomics, genomics, when the collection of the individual data entities brings great additional value. This type of synergy effect is not so noticeable for more artefactual types of experimental data, for example, a set of electrophysiological recordings from the auditory cortex of the rat, or an ecological study of fungal diversity in a semi-arid region.

Because of this there is growing funder/peer pressure to make all types of data set open access and repositories such as Dryad and figshare are designed to meet this need, although their appeal to date has been quite selective.

Aside from extensive data reuse, funders providing incentives that promote data data deposition attribution of data authorship (hopefully using ORCIDs), and research consortia and publishers implementing appropriate global standards and production processes. As discussed, the importance of centralising repositories will vary from data type to data type and the costs of providing these additional services will fall on the funders and publishers.

Social Sciences and Humanities [DANS]

In this section we will discuss data citation practices in social sciences and in humanities. As these disciplines are heterogeneous, we cover them separately.

Social Sciences

Social sciences cover disciplines such as Communication Sciences, Criminology, Pedagogy, Political Sciences, Psychology and Sociology. It also overlaps with other disciplines like History in the case of Oral History. The social sciences use different methodologies and types of data. This report will focus on survey data as a typical data type that is used in the social sciences.

Survey data is typically collected via questionnaires or research instruments that measure concepts using variables, operationalised as question wordings. Comparative research is often carried out over time (longitudinal studies) and across multiple countries. Longitudinal studies can be separated into studies that are repeated over time using the same panel (respondents) in multiple waves, or using the same questions and different respondents (cross section studies). Studies are conducted for scientific research, or for other purposes such as policy making. Data are published and archived by the research institutes themselves, or via national and international data archives.

Background to Social Science infrastructure.

CESSDA (Consortium of European Social Sciences Data Archives) is a consortium of 14 national social science data centers that brings together social science data archives from across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation[1]. It achieves this goal by providing large scale, integrated and sustainable data services to the social sciences.

DANS (Data Archiving and Networked Services) promotes sustainable access to digital research data, focussing on social sciences and humanities. DANS is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). DANS encourages the reuse of research data

[1] http://cessda.net

by researchers or other interested parties. It does so by making research data accessible via:

- EASY is an online archiving system based on Fedora. EASY allows research data to be shared and preserved. Research data can be deposited by individual researchers or organizations, as a set of data files and metadata.
- NESSTAR and SurveyDataNL allow survey data to be published in a structured format of studies, variables and/or questions.Both services make use of the DDI-metadata[2] standard. CESSDA federated the NESSTAR information into the CESSDA Data Catalogue[3].
- DataverseNL where researchers and lecturers can store, share and register research data online, both during research and for up to ten years afterwards. DataverseNL has been built on a scientific platform developed by Harvard University. In the Netherlands it has been developed into a shared service for 11 institutions, managed by DANS.
- NARCIS is the national portal for those looking for scientific information, including (open access) publications from the repositories of all Dutch universities, KNAW, NWO and various scientific institutions and datasets of several data archives. It provides access to descriptions of research projects, publications, experts and research institutes in the Netherlands.

The DANS data archive collection provides access to thousands of datasets in the fields of humanities, archaeology, geospatial sciences and behavioural and social sciences. Contracts have been concluded with various institutions, such as the Netherlands Organisation for Scientific Research (NWO), the Dutch Land Registry Office (Kadaster), Statistics Netherlands (CBS) and the national government, with regard to the delivery of new data for scientific research.

Current status in social sciences - DANS as an example

At DANS, datasets are deposited as collections of files. For datasets based on survey research, these collections usually consist of one or more files with survey responses in SPSS format or another statistical package, codebooks with a description of the fieldwork and other methodological aspects and examples of the survey instrument. For other types of social research the data collections contain other data types, for example audio/visual interviews or images.

DANS long-term archive EASY provides these datasets with generic metadata at a study level. Each dataset will receive a unique DOI (provided by DataCite) when published. The DOI links to the to the information page of the dataset. An instruction to the user of the dataset is provided on the information page, on how to cite the dataset correctly using the DOI.

[2] http://www.ddialliance.org/
[3] http://cessda.net/Data-Catalogue

The predecessors of DANS actively documented references of associated literature to the datasets. But this was not captured in a very standardized way. These collections of datasets are now part of the DANS collection in EASY. These literature references are registered as relations, without identifiers or URIs. Datasets from the last 10 years can have relations with URIs, but these are often not classified. Besides literature references these relations can refer to, for example, a project website or other information sources. DANS currently estimates that there are between 1.000 and 5.000 references hidden in the archive. DANS is planning to transform these hidden references into actionable and standardized references and provide these to DataCite and OpenAIRE.

Advanced CESSDA archives such as GESIS provide documentation on lower-level items: concepts, variables and questions. Datasets are (more often than other types of items) given persistent identifiers. GESIS hosts a DataCite registration agency dedicated to social science datasets[4] and is currently working on the InFoLiS project[5], connecting research data and publications.

CESSDA has no explicit central policy or guideline on data citation. A look at the different data centers show that most of them published terms of use that require data users of data to attribute the data they use in their publications. An explicit requirement to cite the data using persistent identifiers is rarely mentioned.

Publishing perspective

Compared to other disciplines, books and "grey literature" reports are more commonly published outputs than journal articles. Currently few articles published in social science research journals cite the data used in the underlying research.

In the social sciences data sources are sometimes listed in the references, sometimes in the text, or included in tables and footnotes. Only occasionally do these references provide enough information to guarantee future access to the identical data set. Data citation standards are set by professional societies and/or publishers. For example the MIT Libraries advise authors to include the following information when citing datasets: Authors names, Title, year of publication, publisher or distributor, URL, identifier, or other access location. In Endnote authors are advised to use the reference type for "dataset." If using Mendeley or Zotero, they should make do with using other more generic reference type templates and fill in the essentials for the dataset.

Humanities

The humanities cover a wide range of disciplines such as archaeology, history, linguistics, literature or theology. Other than making bibliographic references to sources, data citation is not yet concern for the majority of humanities researchers. We will illustrate data citation in the humanities by providing some characteristics of their (data) sources, as well as trends and examples of data citation.

[4] http://www.da-ra.de/en/home
[5] http://infolis.github.io/about/

The (data) sources used by humanities are heterogeneous and can at best be generalized as a collection of resources or files. The file types vary from text documents (e.g. literature), to audiovisual files (e.g. interviews or movies) to databases (e.g. digitized census data). A archaeological dataset may for example consist of excavation photos, maps of the site, databases on the artefacts and a report. In some cases the data is deposited as a collection of such files which can be deposited in a generic archive such as DANS EASY, in other cases the data is aggregated in a dedicated and more structured collections (such as The Language Archive[6] for linguistic resources).

On a European scale, communities or infrastructures promote best practices, standards and infrastructure to support the description and deposit of such resources. Examples of larger communities or infrastructures are DARIAH (humanities), CLARIN (linguistics), ARIADNE (archaeology) or EHRI (holocaust research). An important concern in such communities is the identification of resources. Initially to ensure long-term access to their resources, but increasingly from the perspective of data citation. However, requirements or recommendations for enabling data identification or data citation take time before they are acknowledged and adopted by the many smaller communities and data providers.

A big community that is important for (and overlaps with) humanities is the cultural heritage community. Museums, libraries and archives provide access to large amounts of cultural heritage resources which are used by humanities researchers. There is awareness and growing support for persistent identification of cultural heritage resources, but support for data citation is less important, as research is not the primary objective of cultural heritage.

Challenges are the availability of identification and description of resources on the granularity that researchers need, and the availability of digital resources. In many cases, resources are not digitized yet, and the metadata is only available on the level of the archive or the collection, instead of on item-level.

Two examples of how data in the humanities can be cited. One generic archive using Datacite DOIs, and one humanities resource without any.

In DANS EASY:

> Dr L.J. Touwen, Universiteit Leiden, Vakgroep Geschiedenis: Shipping and trade in the Java Sea, 1870-1940. DANS. http://dx.doi.org/10.17026/dans-zze-qamz

Van Gogh letters by Huygens ING and Van Gogh Museum[7]:

> Leo Jansen, Hans Luijten, Nienke Bakker (eds.) (2009), *Vincent van Gogh - The Letters*. Version: December 2010. Amsterdam & The Hague: Van Gogh Museum & Huygens ING. http://vangoghletters.org. Consult the homepage for the current version.

[6] https://tla.mpi.nl
[7] http://vangoghletters.org/vg/about_6.html#intro.VI.6.3.

Other important trends that will affect data citation are Digital Humanities[8] and Linked Data[9]. Digital humanities is the area where humanities research is combined with computational methods. This will raise the digital awareness and need for trustworthy data citation. Linked Data is gaining momentum in cultural heritage and humanities as it allows the use and integration of heterogeneous or loosely-structured data and metadata. This will also increase the awareness and demand for online digital resources that need to be referrable.

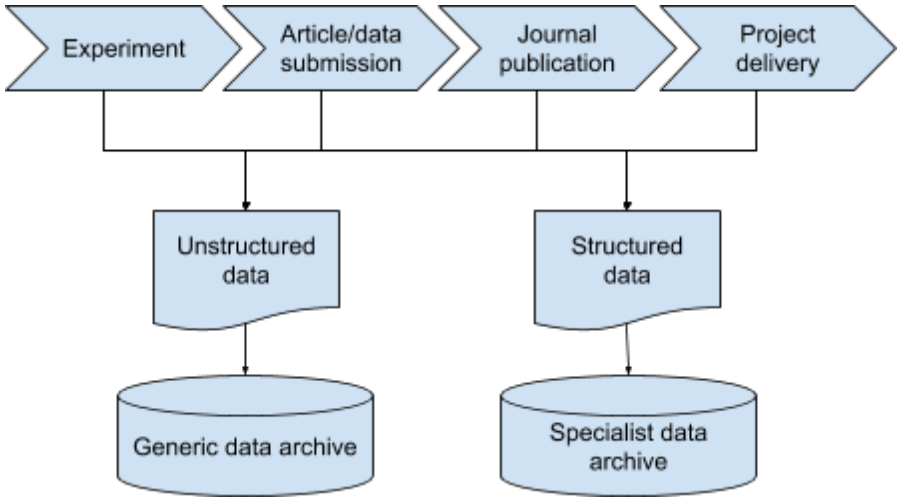
Data in the humanities are heterogeneous, but their online availability is improving, as well as their identification using URIs or even persistent identifiers. DOIs are primarily provided via generic data repositories.

Citation of the data is also not yet common and needs to be promoted. Communities like CLARIN are endorsing the Data Citation principles. DANS is promoting data availability policies with publishers[10] and promoting data as publishable objects via a data journal with Brill Publishers[11] (assignment of DOI is in progress).

Overview and opportunities: a roadmap for OpenAIRE

From our analysis of different patterns of data use and reuse we discovered that the standard article/data publishing workflow has been modified to accommodate the publication of data in a different ways in different fields. (See graphic below). In some instances, e..g. ENA and PDB, the expected behaviour of the research community, supported by publishers, is to deposit data in a recognised publisher-independent repository prior to publication. More recently, the increasingly data-rich papers being published and a drive to improve reproducibility is leading to an increase in the number of general repositories such as PANGAEA, DANS' EASY, Dryad, Zenodo, and Figshare. Indeed, publishers such as Elsevier/Mendeley have recently built their own data repositories.

[8] https://en.wikipedia.org/wiki/Digital_humanities
[9] https://en.wikipedia.org/wiki/Linked_data
[10] see inventory of Leen Breure: http://xposre.nl/cliodap/DAP_Ideal+Practice_1-1.pdf
[11] http://dansdatajournal.nl/



Data is deposited at different phases during research. Depending on the practices and standards in the discipline, these are deposited in generic of specialist data archives.

There were also three distinct views on how closely literature and data should be related during the publishing process, namely: the data is an essential, supplementary component of the article; the article is essentially an annotation of the data; the data constitute an independent published entity. These variants are not discipline-specific - geo- and life sciences recognise all three.

In many cases the data set is enhanced by providing a summary Web page and/or a minimal description of the study as part of the header to the actual data file (e.g. DANS, ENA, PDB). The home page may also be linked to a curated bibliography or to other types of contextual information.

Making the link

Thus, a data set is submitted, verified and allocated a unique identifier such as a DOI or a an accession number, ideally. The former service can be provided by an outside agency such as DataCite, the latter by the repository itself.

How then to create a well-formed, machine readable data citation such as this one in an article:

Cossu F., Milani M., Mastrangelo E., Bolognesi M. (27 Oct 2009). Crystal structure of XIAP-BIR3 in complex with a bivalent compound. PDB 3g76 [http://www.ebi.ac.uk/pdbe/entry/pdb/3g76].

This citation was created using metadata from PDB, applying the Force11 data citation principles [Data Citation Synthesis Group, 2014] to the JATS v1.1 draft, and then transforming the XML to text. For a data-citation the Force11 data citation principles describe characteristics of how data should be cited. The eight principles are:

1. Importance
2. Credit and attribution
3. Evidence

4. Unique identification
5. Access
6. Persistence
7. Specificity and verifiability
8. Interoperability and flexibility

Since the importance of the cited data and the need for evidence are the responsibility of the citing author principles 1 and 3 are not considered here. To fulfill the other principles we attempted to map usually available metadata to the data citation principles. For this purpose the unique identifier/ accession allows unique identification (4) persistence (6) and, for the repositories examined, access (5), as a URL can be constructed knowing the accession.

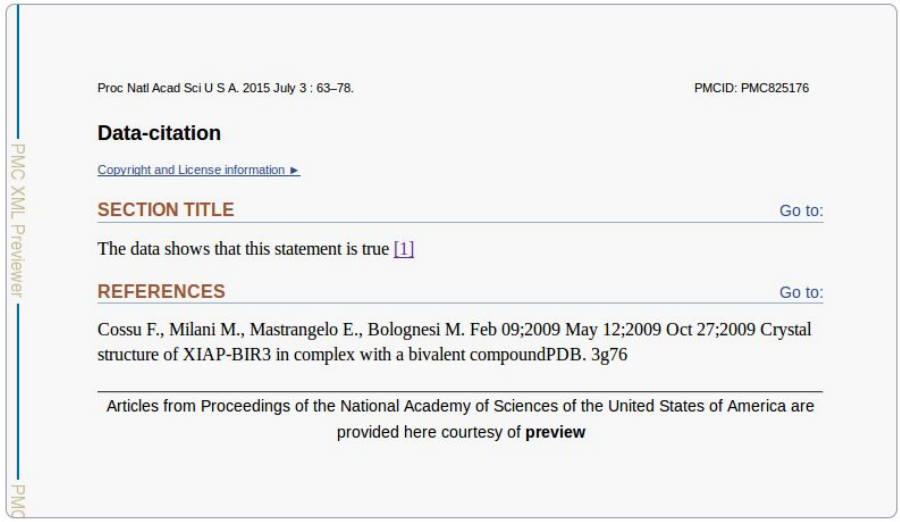
Credit and Attribution are (2) given using submitter metadata. Principle 7, specificity and verifiability, can be satisfied by a version number. This however seems to not always be present but inclusion of a modification date may to a limited degree be a substitute. Interoperability is granted using the JATS format and allows easy machine reading of data citations.

Consequently a JATS v1.1 compliant example XML file was created from metadata retrieved through the PDBe API and is shown below. A guide on using JATS for data citation [Lapeyre, 2015] by staff from the company involved in writing the JATS specification was used as a guide for this prototype. In contrast to the preference expressed by Lapeyre the element-citation was chosen over the mixed-citation as it does not put any constraints on the format of the reference list item and delivers the metadata allowing freedom over the rendered format of the data citation. That is the data citation reference list item shown below was generated transforming the element-citation from below using an XSL transformation.

Example of JATS-compliant data citation XML

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
    <front/>
    <body>
        <sec>
            <p>The data shows that this statement is true <xref ref-type="bibr" rid="ref1"
                >[1]</xref></p>
        </sec>
    </body>
    <back>
        <ref-list>
            <ref id="ref1">
                <element-citation publication-type="data">
                    <person-group person-group-type="submitter">
                        <name>
                            <surname>Cossu</surname>
                            <given-names>F.</given-names>
                        </name>
                        <name> [3 lines]
                        <name> [3 lines]
                        <name> [3 lines]
                    </person-group>
                    <date iso-8601-date="2009-02-09" date-type="received"> [4 lines]
                    <date iso-8601-date="2009-05-12" date-type="pub"> [4 lines]
                    <date iso-8601-date="2009-10-27" date-type="corrected">
                        <year>2009</year>
                        <month>10</month>
                        <day>27</day>
                    </date>
                    <data-title>Crystal structure of XIAP-BIR3 in complex with a bivalent
                        compound</data-title>
                    <source>PDB</source>
                    <pub-id pub-id-type="accession"
                        xlink:href="http://www.ebi.ac.uk/pdbe/entry/pdb/3g76"
                        assigning-authority="protein data bank">3g76</pub-id>
                </element-citation>
            </ref>
        </ref-list>
    </back>
</article>
```

In the screenshot below it can be seen what a minimal example of a data-citing article may look like. In this case it was rendered using the PMC XML Previewer.



Factors influencing machine-readable data citation

Implementing data citation effectively for the different scientific disciplines poses many challenges, but these are chiefly associated with social factors rather than technical. These challenges can be categorized according to the following 'reference workflow':

1. **Researchers need to deposit data in a community-accepted archive** that
2. **Data repository should follow the FAIR data principles (Findable, Accessible, Interoperable and Re-usable), or similar, for data.** Activities needed to provide such an infrastructure can include certification (Data Seal of Approval, WDS, NESTOR) and the process of allocation of persistent identification (Datacite). While some difficulties may remain concerning the handling of some particular types of data (such as clinical data), or data processes such as data fragments, or complex data versioning, many best-practice examples of solutions to these challenges may have been developed or are available in other fields.
3. **Researchers must cite data in their articles**. Many funding agencies and leading STM and SSH publishers are beginning to require data deposition statements and/or data citation. Linking published data sets to individual researcher ORCIDs could help to motivate adoption of this practice if data sets are recognised as legitimate research outputs by funding agencies.
4. **Journals must publish data citations in machine-readable formats**. We show here that JATS is capable of supporting such links when the database supplies the appropriate metadata. This is a by-far preferable route to discovering data citations over post-publication text mining approaches.
5. **Data repositories must link to articles.** This requires the data repositories to inform themselves about citations from journals. Interlinking services like the DLI, Datacite and Crossref promise to support rigorous cross linking. This may require the development of new standards for exchanging such cross links.
6. **Data citations must provide an added value**. Having access to data will improve the peer review process and enhance research reproducibility. Furthemore citing data is an indicator of the re-usefulness of the archive or dataset beyond the study that generated it. Being able to map such impacts is of use not only to the

researchers who generated the dataset(s) but also funders and data infrastructure providers.

What additional discipline-specific issues must be taken into account for future planning?

What we have found is that beyond differences in nomenclature, the different areas are very similar in their approaches to data citation. Problems that remain to be addressed include:

- Forms of licensing
- Personal data, privacy
- Representation of data timeline, versioning

Outcomes and next steps

With the JATS standard allowing machine-readable data citations and a proposed implementation available, relevant players in the field of data citation and publishing (e.g. publishers and data repositories) should be approached and encouraged to implement this JATS data citation practice. Conferences and workshops like the RDA plenary meetings in Tokyo present opportunities for this, as well as the recently held Force 11 Data Citation Implementation Pilot Workshop in Boston, USA (February 3, 2016).

1. Publishers keen to adopt these approaches may find a showcase of tools to assist implementation useful. We plan to build a prototype that, given an accession number, demonstrates an automatically generated JATS element-citation snippet (for machines) and a text data citation like the example above (for the human reader).
2. In OpenAIRE's own infrastructure the data-literature interlinking service (DLI) should be evaluated for the ability to support of JATS as an output format and could hence be turned into a showcase to support adoption of good data citation practices.
3. Ideally PID resolvers should offer common APIs and return common minimal metadata for data citations. We will explore the matter of such standards with organizations such as DataCite, CrossRef and identifiers.org.
4. There are currently other initiatives that have capability for data-literature interlinking services similar to OpenAIRE. We will work with those organizations (which include CrossRef) to build common technical approaches and share information.
5. The development of tools as described in (1) above will provide a focus for discussion across disciplines - informing not only publishers but also data repositories as to how to provide better metadata to support machine-readable data citation.

Appendix: Overview of initiatives and resources around data citation

Communities:

- **Force11** (https://www.force11.org) has been active and completed working groups around data citation. These include working groups for BioCaddie Extension (focus on publishers, data citation principles, software citation, etc).
- **Research Data Alliance** (RDA, https://rd-alliance.org) hosts interest- and working groups focused on data citation. These include working groups for Data Publication, for Data Citation and for Publishing Workflows.
- **Codata/ICSTI** hosts a taskgroup on Data Citation Standards and Practices. http://www.codata.org/task-groups/data-citation-standards-and-practices
- **DataCite** (https://www.datacite.org/about-datacite/working-groups) hosts two working grous: on Metadata and on Policy and Best Practices.
- ?IOC SCOR WG on data citation

Projects

- **InfoLiS** (http://infolis.github.io). The goal of the InFoLiS project is to connect research data and publications. Links between data and literature are created automatically and made available for seamless integration into different retrieval systems.
- **THOR** (http://project-thor.eu). THOR will “establish seamless integration between articles, data, and researchers across the research lifecycle. This will create a wealth of open resources and foster a sustainable international e-infrastructure”.

Services:

- DataCite
- Data Citation Index by Thompson Reuters. http://wokinfo.com//products_tools/multidisciplinary/dci/about/
- CrossRef Events tracker. http://det.labs.crossref.org
- JISC’s Journal Research Data Policy Bank (JoRD). https://jordproject.wordpress.com
- Material:
    - Bibliography on data publishing literature incl. data citation (https://goo.gl/bnkt0x)

Meetings and Events

- Washington DC data citation meeting (Fall 2016)
- BioCaddie Workshops (publishers) Feb 2016
- RDA Tokyo (Spring 2016)

Information:

- ANDS Webinar - Data Citation Series: https://www.youtube.com/playlist?list=PLG25fMbdLRa4peWpeZslW0cLSPYNjcbc1
- DRYAD Data Citation Guidelines. http://wiki.datadryad.org/Citing_Data

Literature:

- [Boulton, G. et al., 2012]
  Science as an Open Enterprise - Final Report.
  https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf
- [Starr et al., 2015]
  Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1 https://dx.doi.org/10.7717/peerj-cs.1. This article contains recommendations of "the **Identifiers, Metadata, and Machine Accessibility** group" (i.e. the authors) on how to achieve human and machine accessibility of cited data. Seeing that several of the authors are in some of the above mentioned working groups, this article represents / connects with very recent thinking about the topic in various WGs.
- [Hoogerwerf et al., 2013]
  Hoogerwerf, Maarten, Mathias Lösch, Jochen Schirrwagen, Sarah Callaghan, Paolo Manghi, Katerina Iatropoulou, Dimitra Keramida, and Najla Rettberg. “Linking Data and Publications: Towards a Cross-Disciplinary Approach.” *International Journal of Digital Curation* 8, no. 1 (June 20, 2013). doi:10.2218/ijdc.v8i1.257.

  This article is an output of OpenAIREplus, which built and evaluated demonstrations of presentations of text publications in their context of data, people and research projects, in two different disciplines. It recommends that the relationships should be captured by "the most knowledgeable stakeholder, usually the author or creator of a resource" [Hoogerwerf et al., 2013] and that methods for automatically finding links afterwards are explored.

- [Armstrong, 2012]
  Armstrong, John. “A Question Universities Need to Answer: Why Do We Research?” *The Conversation*, April 9, 2012.
  https://theconversation.com/a-question-universities-need-to-answer-why-do-we-research-6230.
- [Ball and Duke, 2015]
  Ball, Alex, and Monica Duke. “How to Cite Datasets and Link to Publications.” Digital Curation Centre, July 30, 2015.
  http://www.dcc.ac.uk/resources/how-guides/cite-datasets.

  Overview of techniques and initiatives for interlinking data and publications. The OpenAIRE Data-Literature Interlinking service is described, just like the JDDCP and other things. Originally published in 2011, updated July 2015.
- [Mathiak and Boland, 2015]
  Mathiak, Brigitte, and Katarina Boland. “Challenges in Matching Dataset Citation

Strings to Datasets in Social Science." *D-Lib Magazine* 21, no. 1/2 (January 2015). doi:10.1045/january2015-mathiak.

Update on the work that tries to mine textual references to data and organise these into patterns. The InfoLiS project provides the context for this article.

- [Data Citation Synthesis Group, 2014]
  Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].
- [Lapeyre, 2015]
  Deborah Aleyne Lapeyre. "Citing Data in Journal Articles using JATS" last accessed 15.01.2016
  https://www.force11.org/sites/default/files/d7/project/882/citing-data-in-jats-2015-06.pdf