DATA CITATION STANDARDS AND INDEX REQUIREMENTS

11.04.2016

OpenAIRE2020

Open Access Infrastructure for Research in Europe towards 2020
Deliverable Code: D7.5 - Version (0.2 – Draft)
PUBLIC

This deliverable report gives an overview of data citation practices across earth-, life-, social sciences and humanities. It furthermore introduces a tool developed to ease standard compliant data citation and discusses data citation indexes and metrics.

H2020-EINFRA-2014-1
Topic: e-Infrastructure for Open Access
Research & Innovation action
Grant Agreement 643410

Document Description

D7.5 – Data citation standards and index requirements

WP7 - Scholarly Communications R&D	
WP participating organizations: EMBL, UniHB, DANS, CNR, UoA, UGOE, ICM, UvA, SURF, TU DELFT, COUPERIN	
Contractual Delivery Date: 03/2016	Actual Delivery Date: 10/2016
Nature: Report	Version: Final
Public Deliverable	

Preparation Slip

	Name	Organisation	Date
From	Florian Graef	EMBL-EBI	11/04/2016
	Jo McEntyre	EMBL-EBI	11/04/2016
Edited by	Florian Graef	EMBL-EBI	01/06/2016
	Jo McEntyre	EMBL-EBI	01/06/2016
Reviewed by	Paolo Manghi	CNR	24/06/2016
	Natalia Manola	University of Athens	24/06/2016
Approved by	Tony Ross-Hellauer	UGOE	14/10/2016
For delivery	Mike Chatzopoulos	University of Athens	14/10/2016

Document Change Record

Issue	Item	Reason for Change	Author	Organization
V0.1	Draft	Initial version of Document	Florian Graef	EMBL-EBI
V0.2	Draft	Revision	Florian Graef, Jo McEntyre	EMBL-EBI
V0.3	Draft	Reviewers Feedback	Florian Graef, Johanna McEntyre	EMBL-EBI

Table of Contents

Table of Figures

Disclaimer

This document contains description of the OpenAIRE2020 project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the OpenAIRE2020 consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

OpenAIRE2020is a project funded by the European Union (Grant Agreement No 643410).

Acronyms

JATS	Journal Archiving Tag Suite
RDA	Research Data Alliance
DLI Service	Data Literature Interlinking (Service)
DANS	Data Archiving and Networked Services
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
XSL	Extensible Stylesheet Language
XML	Extensible Markup Language
PID	Persistent Identifier
PMID	PubMed Identifier
DOI	Digital Object Identifier
ENA	European Nucleotide Archive
PDBe	Protein Data Bank Europe
THOR	Technical and Human Infrastructure for Open Research
FORCE11	Future of Research Communications and e-Scholarship
EASY	Electronic Archiving System
ANSI	American National Standards Institute
NISO	National Information Standards Organization
API	Application Programming Interface
URL	Uniform Research Identifier
INSDC	International Nucleotide Sequence Database Collaboration
H2020	Horizon2020

Publishable Summary

Research has become increasingly data intensive through the emergence of new experimental techniques and the development of tools that operate on this data. Furthermore open access policies are making even more data and publications widely available. Data are stored in a plethora of very diverse repositories that can be specific for each data type, community or discipline, as well as in more generalist data repositories. Citing data provides a definitive means for users to navigate this complex space, offering provenance of scientific assertions in articles, supporting the hypothesis the author states through improved verifiability of results. In addition, citing data grants credit to the creator of a dataset more specifically than through citing and entire paper. Finally, large-scale analysis of data reuse contributes to our understanding of the impact of particular areas of research or funding programmes. While mentioning data in narratives may be clear to human readers, to enable sharing and automatic analysis of data-literature cross links, as would be required to construct, for example, data citation indexes, the implementation of standard, machine-readable data citations is necessary.

This work addresses several aspects and requirements to make this vision of full interoperability between the literature and data a reality.

Firstly, we reviewed data citation practices across different disciplines and shared information on both discipline-specific and cross-disciplinary global initiatives in this area. In the life sciences, data is cited relatively frequently, but not in a standard way. In the earth sciences often publications do not cite data but data records cite the publication they belong to. While in the social sciences and humanities, research can heterogeneous and data citation practices exist, but are specific to individual communities or repositories. We identified a standard for article tagging (JATS) that now supports the cross-disciplinary Data Citation Principles that is commonly used in the life sciences but could be applicable across disciplines.

Secondly, in recognition that in the life sciences, data is cited relatively frequently, and that there is now support to do this in a standard way (JATS), we have developed a prototype tool that generates formal data citations, given a data accession number (PID) as input. This tool has been introduced to the journal publishing community through the Force11 Data Citation Implementation Group.

Thirdly, we have explored the challenges of constructing a data citation index in the life sciences through the analysis of large datasets of data citations of two major life sciences data resources, the articles cited by the data records within those resources, and the relative citation of the data themselves versus the articles that describe the data in the first place. This has been a complex project, but has already contributed to a paper on the impact of data resources, assessed via data citations in articles and patents, and furthermore we expect to publish a further article in due course.

Finally, all code, datasets and completed articles have been published on open forums (Zenodo, GitHub, and F1000R). The work described here continues, involving several groups and projects (such as the H2020 project THOR). We plan to present the latest outcomes of this work at a number of upcoming forums, including the RDA Barcelona meeting in April 2017.

1 | INTRODUCTION

1.1 Why cite data?

Research artefacts from all disciplines ideally form a network including, but not limited to: research articles, data and software. Traversing this graph allows relevant knowledge to be found easily and enables findings to be made and/or validated. This graph serves the scientific community by connecting information and knowledge as well as providing the basis for the scientific credit system.

With the emergence of research methods that have a high data output, as well as an increasing number of open access policies in place, the graph is expected to grow. This is demonstrated by the exponential growth of nucleotide data of the European Nucleotide Archive over more than 30 years[1] as well as the continuous increase in registered research repositories in the re3data registry[2]. Citing data effectively, in particular in research articles, serves two purposes: (1) to clearly show the provenance of the scientific assertions made in the article and (2) to credit the data generators effectively. Thus it will become increasingly important to give data (and other research objects) the same status as research articles, which means that minimally they need to be cited clearly and effectively.

Data citation practices differ across the disciplines represented by the OpenAIRE partners of task 7.3, literature-data integration, i.e., earth and environmental sciences, life sciences and social sciences and humanities. However, data citation practices in general have, until recently, lacked the rigour that has become standard when citing articles. Typically, data is simply mentioned in the flow of a narrative (for example, referring to the identifier and repository) and even when the mode of citation is more formal, the citation may only be of use to a human reader. However, in order to generate a fully interconnected network of research objects, the data citation should also be machine readable, which means that the citations need also to be structured in standard ways. The interconnected multi hub model, laid out in OpenAIRE deliverable 7.3: Governance and Requirements of Interlinking Service, aims to serve this purpose. The multi hub model is briefly outlined below in 2.2.2 Interoperability and sharing of data citations.

There are currently many ongoing efforts across disciplines aiming to improve the manner in which data is cited. The major ones are described further in section 2.1.1 *Existing Initiatives*. These efforts include publishers and data providers in particular.

1.2 The Data Citation Principles

FORCE11, one of several organizations that has been promoting data citation, coordinated the formation of the Joint Declaration of Data Citation Principles[3] which is now endorsed by more than 100 organizations and 240 individuals worldwide. The principles serve as a guideline to build tools and practices on.

The eight data citation principles are[1]:

1. **Importance**
   Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. **Credit and attribution**
   Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. **Evidence**
   In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. **Unique identification**
   A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. **Access**
   Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. **Persistence**
   Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.
7. **Specificity and verifiability**
   Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
8. **Interoperability and flexibility**
   Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

1.3 Implementing the Data Citation Principles and implications for a Data Citation Index in the context of OpenAIRE

OpenAIRE began as support mechanism for the European Commissions' FP7 pilot and Horizon2020 open access policies to collect H2020 research output. It grew into a network of open access advocates collaborating to aggregate and interlink entities of research like publications, data, funding, people and organizations. It enables users to navigate through these information and provides services on top of them ranging from deposition to statistics[4].

In Task 7.3, Literature-Data Interlinking, the aim is to explore and improve the integration of literature and data.

[1] Retrieved from https://www.force11.org/group/joint-declaration-data-citation-principles-final

Aware of this context we identified these goals:

1. Understand current data citation practices across different disciplines
2. Share information on standards, processes, tools and resources that support data citation according to the data citation principles.
3. Where possible, develop tools and materials to encourage wider data citation based on existing tools and encourage adoption
4. Understand requirements for a possible data citation index, developing a prototype and exploring implications.

2 | RESULTS

2.1 Implementing the Data Citation Principles: state of data citation across different disciplines

2.1.1 Existing Initiatives

The following is a summary of the mini-workshop we held to understand the challenges of data citation across disciplines. The full write up of the workshop can be found here[5].

There are many initiatives aimed at improving data citation practices and encourage greater adoption of data citation by scientists and journals. A stakeholder analysis carried out by DANS[6] influenced the joint document of OpenAIRE2020 Task 7.3[5]. Among other information it identified key data citation initiatives:

- ☐ OpenAIRE[4]: OpenAIRE began as support mechanism for the European Commissions' FP7 pilot and Horizon2020 open access policies to collect H2020 research output. It grew into a network of open access advocates collaborating to aggregate and interlink entities of research like publications, data, funding, people and organizations. It enables users to navigate through these information and provides services on top of them ranging from deposition to statistics.

- ☐ FORCE11[7]: Data Citation Implementation Group, Data Citation Implementation Pilot. The most notable outcome are perhaps the Joint Declaration of Data Citation Principles[3]. Furthermore the data citation implementation group formulated the recommendation for the data citation extension of JATS, which was implemented in the latest version of the tag suite.

- ☐ Research Data Alliance[8]: Data Citation Working Group. This WG has compiled a set of data citation recommendations and guides data centres adopting the guidelines. Generally the RDAs goal is to enable data sharing globally across "domain, research, national, geographical and generational boundaries"[9].

- ☐ DataCite: Metadata Working Group, Policy and Best Practices Working Group[10]. DataCite aims to support discovery and citation of data, provide persistent identifiers for data and enable linking of research articles. To achieve that DataCite is working on data citation standards and host a metadata store and provide persistent identifiers (DOIs) for data providers.

- ☐ InFoLis: Integration of Research, Literature and Data[11]. An initiative from the social sciences provides infrastructure and algorithms to find links between publications and research data and attempts to improve reusability of the generated links.

2.1.2 Citing Data – Disciplinary Considerations

The question for the partners in this task was therefore how best to coordinate and contribute with these global efforts from the OpenAIRE project. It soon became apparent that task partners knew little of the current status of data citation in disciplines other than their own, and that while the basic principles of citing data were transferable, there were different contextual challenges

(described below) originating in the norms of prevailing academic practices in different areas of scholarship.

We set out to take a snapshot of the data citation practices in earth, life, social sciences and the humanities to gain a good understanding of the differences and commonalities in the way data is cited, as well as in the respective priorities to improve data citation practices. To do this we conducted a mini-workshop across the task participants, which resulted in a joint document[5] summarizing the data citation practices. It took place virtually on the 11th of November 2015 with project partners Michael Diepenbroek (PANGAEA), Paolo Manghi (CNR), Uwe Schindler (PANGAEA), Marten Hoogerwerf (DANS), Jo McEntyre (EMBL–EBI) and Florian Graef (EMBL-EBI). It served as a knowledge exchange about data citation across the disciplines, because all the representatives operate in key data resources that coordinate with a variety of journals. From our discussions, it is clear that approaches to data citation differ across different areas of scholarship, depending on the existing data infrastructures and difference disciplinary norms. The findings of the document[5] are summed up below.

LIFE SCIENCES

In the life sciences the publishing author is required by journals to deposit certain types of data like DNA sequences, protein structures and gene-expression assays in appropriate data resources. In these resources the data record can be validated and curated and is assigned an accession number/PID (workflows vary). The PID is typically used to cite data in three ways:

1. structured reference in text or supplemental material
2. Data cited (structured) in reference list
3. Unstructured mentions in text or supplemental material

We found that data in the life sciences is commonly cited but this is frequently not done in a structured and machine-readable way but just semantically. E.g.:

"The crystal structure of tyrosinase from ***Agaricus bisporus*** (***Ab***TYR; PDB code, 2Y9X) was chosen as the protein model for the present study."[12]

This represents case 3 and is unfortunately still a very common way of citing data.

Many repositories in the life sciences, like PDB, do have a simple data structure with individual records being contextually mostly independent from other records. However there are more complex ones too. ENA, as a sequence archive, is a much more complex case due to the possible hierarchical organisation of sequences into genomes, chromosomes, genes and even assemblies of reads and individual reads from a sequencing machine.

EARTH SCIENCES

In PANGAEA no restrictions are applied on data submission types but samples taken or measurements made on earth geolocation data is required. Most of the data records are available under creative commons license except from some records of ongoing project but the description and principle investigator, who may act as a gatekeeper for restricted access, are always visible. The data model of PANGAEA follows the standard hierarchy of activities in geoscience data collection. That means that a project consists of expeditions which contain samples or measurements. These samples can be sub-sampled and analyzed with samples

organized in individual datasets. Along with the sample data each dataset contains a metadata record consisting of authors, publication year, title of the dataset, source institution and a DOI.

Three types of citable datasets are supported by PANGAEA. Data supplements which are integral to a scientific paper and its peer-review, data publications that are not directly linked (published through PANGAEA) and peer-reviewed data publications through data journals like *Earth System Science Data*, *Scientific Data* and *Geoscience Data Journal*.

Since in earth science publications data is not frequently cited, PANGAEA keeps track of the "reverse-links" from data to publications. The data literature interlinking service[13] developed as part of task 7.3 is thus an important tool for data repositories to collect information about data-literature (and data-data) links and annotate the publication metadata.[5]

SOCIAL SCIENCES AND HUMANITIES

In the Social sciences and humanities, represented by DANS (Data Archiving and Networked Services), research data can be very heterogeneous. One typical data type that is relatively well structured is survey data. In the case of the DANS data archive survey data is stored in collections of one or more files with survey responses in the format of a statistical package like SPSS alongside descriptions of the fieldwork and other methodological aspects. As a general archive DANS EASY stores other types like audio or video recordings as well as images. At the level of the study general metadata is provided and a DOI minted through DataCite. In contrast to the structured survey data the data sources in the humanities is often heterogeneous and hard to generalize for specialized submission databases. The data of a study range from text documents, audio/video records (interviews and movies) to databases. This data collection of a study is sometimes deposited in a generic archive like DANS EASY[5].

Large communities in Europe are aware of the use of identifiers for resources to ensure both access to the data but also for data citation purposes. However, standards and recommendations take a long time to become widely adopted and applied across all the required stakeholder groups, which includes researchers, journal editors and data resources.

2.1.3 Structured Data Citation with JATS

In all disciplines citing data comes with challenges. As yet there is no cross-disciplinary solution to implementing the Force 11 Data Citation principles universally; and even with technical solutions in place, many of the challenges are social. However, using common formats and standards for citing data would support the development of tools to (a) enable data citation and (b) aid machine readability of scientific publications post-publication. One possible approach could take a lead from the life sciences. A recent revision to the JATS XML standard for describing research articles (a NISO standard already widely used by many journals and the PMC International archive) added support for data citations. This means that data cited in the references lists of articles (according to the Force11 Data Citation Principles) can now be represented in XML tags in the same way as citations to articles, providing not just human but also machine readability of data citations.

However, while JATS is relatively widely used in the life sciences, it is not used across other disciplines. There is no barrier in principle: JATS can support any article type regardless of content matter. Furthermore, the elements required for data citation in JATS can be mapped to

other DTDs used by publishers. However, in practice, publishers have established workflows that are difficult or costly to change that may not involve the use of JATS. Even when JATS is the standard used, there is a time-lag to shifting a workflow to the latest version of JATS, as well as work to then do to make use of the support for data citation. Indeed, the FORCE11 Data Citation Implementation Pilot, described above, is tasked with facilitating pilots to drive structured data citation among early adopters. In addition, research data is highly heterogeneous across the disciplines and the many different resources that archive these data need to also offer the metadata required for citation, as recommended by the data citation principles.

2.1.4 Next Steps

For data citation to become the cultural norm, both journals and data repositories need to change the way they operate, supporting scientists to change the way they cite data in research articles. This will not happen overnight and tools to support this behaviour will be required. For example, data repositories need to offer the full set of citation metadata programmatically, or on web pages, for consumption by reference manager software on a scientist's desktop, that can in turn be used to insert data citations into reference lists when writing an article, which, when published in a journal, can be represented effectively in the XML. In the prototype development described below, we have focussed on the JATS format used by many life science journals, because this is an open standard. As data cited in research articles gets treated more like article citations, this will provide incentives to data repositories to improve their metadata services, as well as incentivising scientists to deposit and cite data more effectively.

2.2 Enabling Data Citation: A tool for generating JATS-compliant Data Citations

In the life sciences the JATS standard (ANSI/NISO Z39.96-2015)[14] is already used for tagging of scientific publications and is capable of supporting data citation. Therefore the EMBL-EBI investigated whether the public APIs of well-known life science databases could be used to retrieve the necessary metadata to support data citation according to the Data Citation Principles, convert the appropriate fields to JATS-compliant XML, and then convert the XML to a text data citation for human readers. While the JATS standard has its roots in the life sciences, it could equally be used in other disciplines and serves as a model for any XML-based article publishing workflow.

The input for the tool (http://www.ebi.ac.uk/europepmc/dcsnippet) is an Accession number (PID) (e.g., AACH01000026 from ENA) from either the European Nucleotide Archive (ENA) or Protein Data Bank (Europe) (PDBe) – two of the most widely used and cited data resources in the life sciences. In the full text corpus of Europe PMC over the past five years (about 1.44 million articles), ENA Accession numbers can be found in about 37,000 articles, and PDB Accession numbers in about 31,000 articles.

2.2.1 Mapping metadata from APIs to the Data Citation Principles

To determine how well the data citation principles could be fulfilled using the metadata available through public APIs. The metadata they provide was aligned to the data citation principles as shown in Table 1. Hereby it is not necessary to map principles one (importance) and

three (evidence) as they are the citing persons responsibility and interoperability and flexibility (principle 8) are given through a common set of metadata fields and formats like the JATS format.

TABLE 1: MAPPING METADATA FIELDS TO DATA CITATION PRINCIPLES

Metadata field	Data Citation Principle
Unique identifier/accession	Unique identification(4), Persistence(6) and Access(5) (with limitations)
Submitter information	Credit and Attribution(2)
Version	Specificity and Verifiability (7)
Deposition date, modification date, publication date	Specificity and Verifiability (7) (with limitations)

The ENA/PDB accessions uniquely identify the record persisted in the database. The access is however not given by the accession on its own but always requires knowledge of the service's canonical URL. For instance, to arrive at the landing page of the ENA record AACH01000026 the URL follows the pattern **http://www.ebi.ac.uk/ena/data/view/{accession}**.

The EBI service identifiers.org[17,18] attempts to solve the problem to resolve to hundreds of life science databases with different accession types providing a stable URL pattern. Identifiers.org uses the pattern

http://identifiers.org/{identifier_key}/{identifier}

to provide a stable URL. The identifier key selects the data collection the identifier is belongs to. This can be one out of over 500 data collections in the MIRIAM registry[17]. Many databases are part of a global network and offer the same content through different websites. E.g. ENA is part of the INSDC (International Nucleotide Sequence Database Collaboration) and presents as well data originated from collaborators like GenBank. Identifiers.org takes this into consideration offering alternative locations to view content.

The (data) submitter information (submitter names) serves the purpose to attribute the effort behind creating the data to the submitter and grant credit. The accession and submitter information was expected to be widely available through the ENA and PDBe services. The specificity and verifiability (7) principle was expected to be a bit more problematic to satisfy as the PDBe API does not support versions of records. Hence deposition date, modification date and publication date were retrieved and included in the JATS format to provide optimum verifiability.

Using text mined accessions of the Europe PMC text mining pipeline from the open access corpus this mapping was used to evaluate to what extent the, through public APIs, available metadata allows citing data according to the data citation principles where text mining just revealed the mention of data in narrative.

In case of PDBe 16,485 records were retrieved. All of them returned an accession number, the title of the data set, the submitter information and submission date, and almost all records (98.5%) returned modification and release dates. No versions were supported.

The same analysis for 164,287 records of ENA data showed that versions were not problematic and every record provided version numbering. Accessions, versions and submission-, modification- and release dates always were present. However, submitter information was only retrieved for 88% of records. This may be due to the complexity of the data records. The submitter names were retrieved from a single reference element in their data schema. Further submitter information could be stored in less obvious field of their data schema.

TABLE 2: METADATA AVAILABILITY THROUGH PUBLIC APIS

	Availability [%]	
Metadata field	ENA	PDBe
Accession	100	100
Submitter name	87.8	100
Version	100	0
Submission date	88.9	100
Modification date	99.9	100
Release date	99.9	100

The lack of some submission dates can be overlooked since it is primarily considered to provide some version specificity in the absence of a version number. Overall the available ENA metadata can be considered well suitable to fulfil the data citation principles.

The PDBe metadata however lacks version information. The modification date can be seen as an inferior substitute and does not allow the same specificity as the version number when citing data.

3.2.2 Tool development

Once we mapped the API outputs to elements required for data citation according to the Data Citation Principles, we generated a corresponding data citation record in JATS v1.1 format. Depicted in Figure 1 is an example for a PDBe record, the element-citation in the minimal example of an article 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>
```

FIGURE 1: DATA CITATION IN JATS V1.1 XML FORMAT

This XML can then be transformed into text for human readers with an XSL transformation:

> *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].*

These two techniques have been embedded in a small web application shown in Figure 2. As input, the tool requires the identifier of a data record and the type of the provided identifier, which typically corresponds with the database.

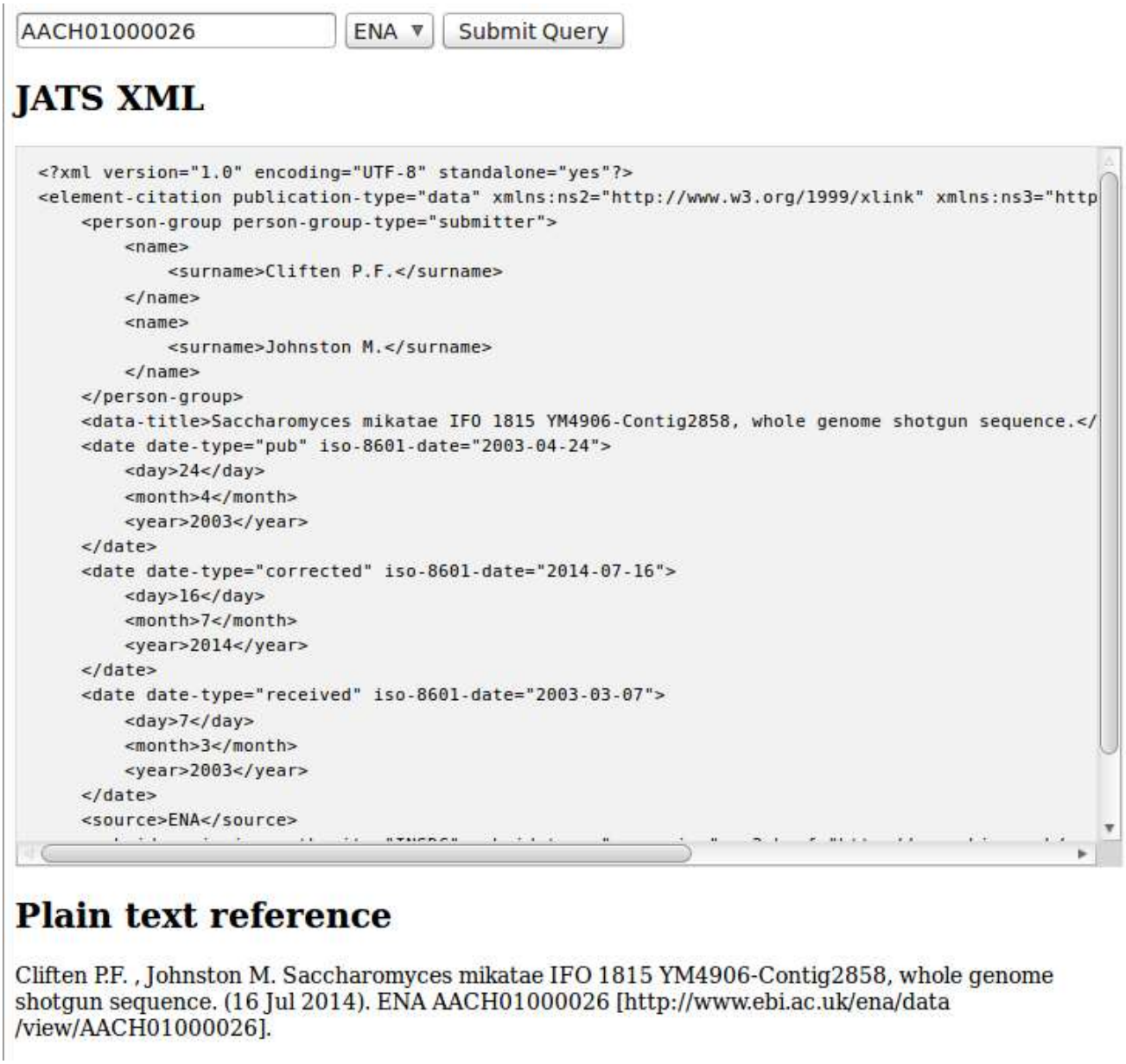

FIGURE 2: OUTPUT OF THE DATA CITATION TOOL

The source code of the tool is open source (Github: https://github.com/FlorianGraef/acc2jats) and we encourage the stakeholders, especially databases and publishers, to contribute to the development, adding support for more databases, reuse the code for own projects and provide feedback.

2.2.2 Interoperability and sharing of data citations

Ideally, with the help of initiatives such as those described above, data citations will in the future be machine-readable. The question then becomes: how would these citations be shared across

infrastructures and made available to others, in order to integrate the literature and data more effectively and support the development of applications based on data citation?

Project partners from PANGAEA and CNR have led efforts in the context of the RDA regarding a data-literature interlinking service (http://dliservice.research-infrastructures.eu/#/). As part of this effort, two workshops were conducted, one of them co-located at RDA Paris the other in Amsterdam, centered around a hub infrastructure for the interlinking service. Participants envision a framework of multiple hubs interfacing different communities as the foundation for a data-literature (and data-data) interlinking service. To replace many bilateral data exchanges the hubs would coordinate the data exchange between hubs. In this context they see CrossRef, DataCite and OpenAIRE as natural hubs for the respective communities of journals, data centres and repositories. This is described in more detail in the deliverable report "D7.3 – Governance and Requirements of Interlinking service"[19] which summarizes the workshops.
The envisioned framework for hub to hub exchange of scholarly links was recently formalized as the Scholix framework with the DLI Service being its first aggregation. It is an outcome of the RDA/WDS Publishing Data Services working group[20,21].

The DLI has already been made available in beta and, among others, EuropePMC contributed data citation links from publications to data records in the European Nucleotide Archive (ENA).[13] In order for such an infrastructure to work, it will be extremely important that there is full reciprocity of contribution across all partners.

2.2.3 Uptake

While the prototype tool described here only operates on two life science data resources, it can be used to show the potential of supporting data citation to both journals and data repositories. For this purpose we are participating in the FORCE11 Data Citation Implementation group and presented our work at F11 DCIP Boston[15]. The aim of the workshop was to gather stakeholders interested in data citation to identify early adopters, and align and harmonize efforts to improve data citation practices and support adoption. We presented our work on using JATS for citing data and announced our data citation tool as part of the early adopters' session. There was a follow-up meeting on this work in London, July 2016, to clarify where publishers stood with respect to implementing data citation.

The tool can be accessed under http://www.ebi.ac.uk/europepmc/dcsnippet and the code is available on Github[16]. Through the FORCE11 DCIP early adopters the company doing JATS conversion for eLife got into contact and intend to develop the data citation prototype tool to support their conversion processes but as well contribute to the open development.

2.2.4 Future work and challenges

Adoption in JATS publishing workflows. We will engage publishers to further test the tool prototype tool to understand how it could work within coordinated data and article publishing

workflows. We will do this via the ongoing FORCE11 workshops, but also in collaboration with the closely related THOR project. We expect to organize an OpenAIRE-THOR joint, cross-disciplinary workshop on data citation within integrated data and publishing workflows in the next 6-12 months.

Adoption in non-JATS publishing workflows. The prototype tool makes it easy to link data correctly from publications. However not all publishers use JATS in their publishing workflows. As the code for this tool is open source, it could be extended or modified to support different publishing formats.

Challenges of scale. The prototype tool operates via the APIs of two life science data resources; scaling this approach to the several hundred life science data resources and more resources from other disciplines that could potentially be cited will be a challenge, as will the ongoing maintenance of the tool in this case. It is possible that collaborating with a resolving aggregator such as identifiers.org may assist in the management of which resources could be included. But this is a serious consideration for any tool operating with this architecture.

An alternative solution could be to encourage data resources to implement schema.org in their web pages and build a tool(s) that operates on those tags. This approach has the advantage of putting the data resources in control of the information they expose for citation, and, assuming that standard operating practices could be agreed, would make the capture of citation metadata from web pages via simple “cite me” buttons or reference manager web browser plug-ins very simple. Furthermore, using schema.org tags in data resource web pages has the added possibility of improved indexing in major search engines such as Google or Bing. Indeed, work is ongoing to investigate the mapping of JATS to schema.org, in particular in the context of ELIXIR[22], so this approach looks promising so far (https://www.elixir-europe.org/news/elixir-interoperability-platform-discusses-google-how-describe-and-share-datasets).

3 | DATA CITATION INDICES

3.1 Overview of Approach

We generated several large datasets of literature-data crosslinks in order to explore the relationships between publication and citation practices around data and articles, focusing on the European Nucleotide Archive and Protein Data Bank. Ideally, the following pattern of behaviour would be expected:

1. Data is deposited, Accession number (PID) provided
2. Article is published, PID is cited in article
3. Data record is updated with article PMID or DOI

This would result in a citation "virtuous circle". If datasets or articles get cited in the future (i.e. reused in some way), then we would expect to see the Accession number or PMID/DOI appear in an article.

The datasets we generated were based on text-mined European Nucleotide Archive and Protein Data Bank Accession numbers (PIDs) from Europe PMC's open access set (1.4 million articles), which, until the Data Citation Principles have been broadly implemented are the proxy for data citations in articles. We also got the articles cited in data records, with all corresponding publication dates, in order to explore the virtuous circle above, and computed the citation counts for all the Accession numbers and PMID/DOIs, to explore their impact, relationships and the construction of a potential data citation index.

3.2 Outcomes

These raw datasets, and a description of their attributes are publically available as dataset 1 in Bousfield et al. "Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources"[23] as files that can easily be imported into spreadsheet software and, after processing, sorted by citation count or any other value.

While simplistically, these data "citations" can be ranked by "number of cites", in practice, we found that there are a number of complexities around the construction of a potential data citation index that operates across all data (or even, just within the life sciences). Not least that (a) the citation of data is still very patchy, citations are scant and approaches are varied (in 2014 2.2/1.8% of publications in the Europe PMC open access set cite data from ENA and PDBe); (b) the virtuous circle described above cannot always be detected, for a number of reasons, for example, the data set does not cite the article, or the mention of the data in the article can not be detected with text mining, (c) different data resources are used in quite different ways, and that the impact of those datasets and resources would be underestimated if citation counts in articles were the only measure. For example, data in archival databases is reused to build "added value" data resources; the primary use of a database may be through website users, or through programmatic access for computation (especially true of big data).

3.3 Future Work

There is currently no way to distinguish between submission citations and reuse citations, which may be useful to know to contribute to assessing the impact of a data resource. While almost all articles (in the hard sciences) generate new data, studies may also use pre-existing datasets that have been reanalysed or have been used for comparative or other purposes. While multiple citations of the same dataset may proxy for reuse, this may not always be the case, as several papers can be based on a single large dataset for the same set of authors. Understanding how and when data are cited and reused and how this may differ from citing articles will be critical to understanding data infrastructure requirements for the future.

Finally, we plan to report on the outcomes of this Deliverable at a forthcoming multidisciplinary meeting, most likely RDA Barcelona 2017 meeting.

4 | REFERENCES

1. Statistics < About the European Nucleotide Archive < European Nucleotide Archive < EMBL-EBI. Available at: http://www.ebi.ac.uk/ena/about/statistics. (Accessed: 20th July 2016)
2. Team, R. org. re3data.org Reaches a Milestone & Begins Offering Badges | re3data.org.
3. Joint Declaration of Data Citation Principles - FINAL. *FORCE11* (2013). Available at: https://www.force11.org/group/joint-declaration-data-citation-principles-final. (Accessed: 10th May 2016)
4. OpenAIRE - OpenAIRE. Available at: https://www.openaire.eu/. (Accessed: 21st September 2016)
5. Bousfield, D. *et al. Data citation practices across earth, life, social sciences and humanities - opportunities for OpenAIRE*. (2016).
6. Hoogerwerf, M. & Companjen, B. Data Citation Stakeholder Analysis.
7. FORCE11. *FORCE11* Available at: https://www.force11.org/. (Accessed: 24th May 2016)
8. RDA | Research Data Sharing without barriers. Available at: https://rd-alliance.org/. (Accessed: 24th May 2016)
9. About RDA. *RDA* (2016). Available at: https://rd-alliance.org/about-rda. (Accessed: 24th May 2016)
10. DataCite | Helping you to find, access and reuse data. Available at: https://www.datacite.org/. (Accessed: 24th May 2016)
11. InFoLiS. Available at: http://infolis.github.io/. (Accessed: 24th May 2016)
12. A, A., H, S., M, P., P, Y. & F, A. In vitro and in silico studies of the inhibitory effects of some novel kojic acid derivatives on tyrosinase enzyme., In vitro and in silico studies of the

inhibitory effects of some novel kojic acid derivatives on tyrosinase enzyme. *Iran. J. Basic Med. Sci. Iran. J. Basic Med. Sci.* **19**, **19**, 132, 132–144 (2016).

13. Data Literature Interlinking Service. Available at: http://dliservice.research-infrastructures.eu/index.html#/. (Accessed: 13th May 2016)
14. Standardized Markup for Journal Articles - National Information Standards Organization. Available at: http://www.niso.org/workrooms/journalmarkup. (Accessed: 24th May 2016)
15. DCIP Boston Workshop Audio Recordings and Slides. *FORCE11* (2016). Available at: https://www.force11.org/group/data-citation-implementation-pilot-dcip/dcip-boston-workshop-audio-recordings-and-slides. (Accessed: 23rd May 2016)
16. Graef, F. FlorianGraef/acc2jats: A proof of concept implementation of the F11 DCP in JATS v1.1. Available at: https://github.com/FlorianGraef/acc2jats. (Accessed: 23rd May 2016)
17. Juty, N., Le Novère, N. & Laibe, C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.* **40**, D580-586 (2012).
18. Identifiers.org < EMBL-EBI. Available at: http://identifiers.org/. (Accessed: 25th May 2016)
19. Michael, D. & Markus, S. OpenAIRE2020-D7-3-GovAndReq-Interlinking-Service.
20. SCHOLIX. Available at: http://www.scholix.org/home. (Accessed: 15th July 2016)
21. RDA/WDS Publishing Data Services WG. *RDA* (2014). Available at: https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html. (Accessed: 15th July 2016)
22. ELIXIR Data for life. Available at: https://www.elixir-europe.org/. (Accessed: 1st June 2016)
23. Bousfield, D. *et al.* Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. *F1000Research* **5**, 160 (2016).