

March 12, 2020

Elena Giglia, OPERAS

Timea Biro, ALLEA E-Humanities Working Group

Humanities and data: for a community driven path towards FAIRness

The road ahead today



1. ALLEA E-HUMANITIES Working Group and Co-OPERAS: what we are doing for you
2. We can give you only hints, as the work is in progress
3. We'll compare the parallel job we did
4. Recommendations match the needs, now...
5. let's build together the tools the community deserves

Who are we?



ALLEA



ALLEA Working Group E-Humanities

The **E-Humanities Working Group** is charged with identifying and raising awareness for priorities and concerns of the Digital Humanities, and contributing to the Open Science and Open Access agenda from a Humanities and Social Sciences perspective, and building consensus for common standards and best practices in E-Humanities scholarship and digitisation.

- **Natalie Harrower** (Chair) – Royal Irish Academy
- **Beat Immenhauser** – Swiss Academies of Arts and Sciences
- **Gerhard Lauer** – Chair of Digital Humanities, University of Basel (Special Member)
- **Maciej Maryl** – Institute of Literary Research of the Polish Academy of Sciences
- **Tito Orlandi** – The National Academy of the Lincei
- **Bernard Rentier** – The Royal Academies for Science and the Arts of Belgium
- **Eveline Wandl-Vogt** – Austrian Academy of Sciences
- **Timea Biro** (Secretariat) – Royal Irish Academy



Main purpose and objectives

The main purpose of the CO-OPERAS IN is the FAIRification of the research process and resources in the SSH, leveraging both on building services, sharing standards and on changing the communication culture in SSH. A second purpose is the contribution of CO-OPERAS network to the FAIR standards from the SSH data.

To improve **Findability**, CO-OPERAS IN will implement

- **Identification services**, already tested in the HIRMEOS project
- **Metadata enrichment**, with data publications crosslinking. It will promote dialogue within

THE RESEARCH INFRASTRUCTURE FOR THE HUMANITIES AND SOCIAL SCIENCES

- coordinating and federating resources
 - nurturing the players
 - taking care of the whole cycle
- TO INCREASE THE OVERALL QUALITY OF THE ECOSYSTEM AND PROVIDE THE SERVICES RESEARCHERS NEED/DESERVE



...working for you

Sustainable and FAIR Data Sharing in the Humanities

ALLEA Report | February 2020

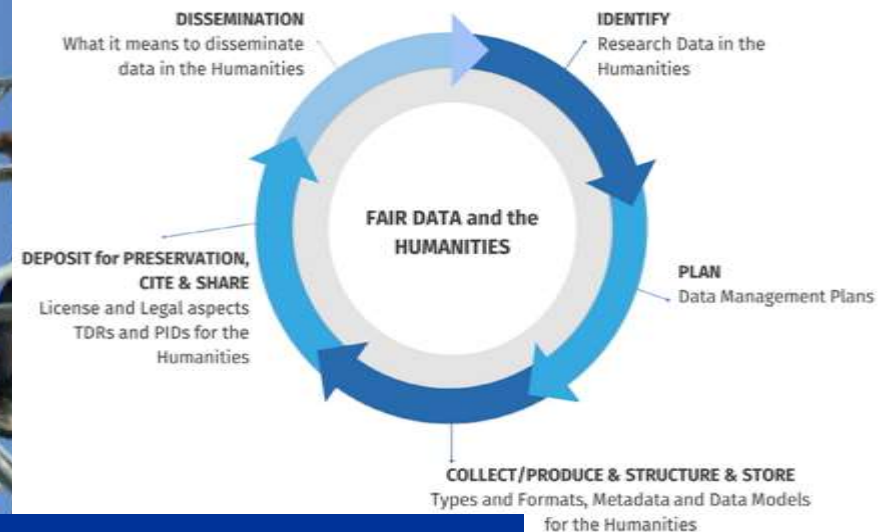
bit.ly/ALLEAFAIR

- the Working Group drafted Recommendations in early 2019
- these were released for an open consultation in the summer of 2019
- a workshop held at DARIAH annual event

200+

The open consultation received over 200 comments and editing suggestions, which were each carefully considered, and used to develop the final version of the recommendations.

- **thank you** to all who participated
- wg carefully considered all comments in creating final draft
- challenging in the possible best way: principles VS concrete; diversity of humanities; FAIR practices is evolving; member state differences



- MAPPED TO THE PHASES OF THE DATA MANAGEMENT LIFECYCLE
- A STARTING POINT TO «CREATIVELY ENGAGE WITH THIS CHANGE»

...working for you / 2



CO-OPERAS

LISTENING TO
THE
COMMUNITY



- 4 NATIONAL WORKSHOPS
[TORINO, COIMBRA, PARIS, GOETTINGEN]
GATHERING 76 RESEARCHERS FROM 32 SSH
DISCIPLINES
- + 1 INTERNATIONAL IN BRUSSELS (9 RES.)
- WHAT ARE «DATA»
- HOW FAIR ARE THE DATA YOU USE/PRODUCE?

...FAIR?

A TRUSTED
REPOSITORIES,
FORMATS

R LICENSES AND
DOCUMENTATION

F METADATA,
PERSISTENT
IDENTIFIERS...

I ONTOLOGIES,
STANDARDS

TO KNOW MORE

Comment | [OPEN](#)

The FAIR Guiding Principles for scientific data management and stewardship

by D. Wilkinson, Michel Dumortier | | Barndt Mann

[FAIR principles](#), Nature, March 2016



It is clear that at the current moment – to say the least – the FAIR principles will persist in shaping the management and sharing of research data for some time.

...why should we care about FAIR data?

The Vienna Declaration on the European Open Science Cloud

Vienna, 23 November 2018

e 2 0
u 1 8
- a t

Vienna, Nov.23, 2018

We, Ministers, delegates and other participants attending the launch event of the European Open Science Cloud (EOSC):

- 1. Recall** the challenges of data driven research in pursuing excellent science as stated in the “EOSC Declaration” signed in Brussels on 10 July 2017.
- 2. Reaffirm** the potential of the European Open Science Cloud to transform the research landscape in Europe. Confirm that the vision of the European Open Science Cloud is that of a research data commons, inclusive of all disciplines and Member States, sustainable in the long-term.
- 3. Recognise** that the implementation of the European Open Science Cloud is a process, not a project, by its nature iterative and based on constant learning and mutual alignment. Highlight the need for continuous dialogue to build trust and consensus among scientists, researchers, funders, users and service providers.
- 4. Highlight** that Europe is well placed to take a global leadership position in the development and application of cloud services for Science. Reaching out over time to and open to the world, reaching out over time to
- 5. Recall** that the Council

SEAMLESS ACCESS TO OPEN BY DEFAULT
FAIR DATA

9. Call for the European Open Science Cloud to provide all researchers in Europe with seamless access to an open-by-default, efficient and cross-disciplinary environment for storing, accessing, reusing and processing research data supported by FAIR data principles.

Science Cloud a reality, hinting at the need to further strengthen the ongoing dialogue across institutions and with stakeholders, for a new governance framework to be launched in Vienna, on 23 November 2018.

EOSC and FAIR



- FAIR (Findable Accessible Interoperable Reusable) is here to stay July 2017
- Turning FAIR into Reality recommend disciplinary approaches to FAIR
- Humanities has particular considerations, unique traditions in approaching the idea of 'data'

- CULTURAL CHANGE: COMMON CULTURE OF DATA STEWARDSHIP
- NO DISCIPLINE, INSTITUTION OR COUNTRY MUST BE LEFT BEHIND

Data culture and FAIR data

- [Data culture] European science must be grounded in a common culture of data stewardship, so that research data is recognised as a significant output of research and is appropriately curated throughout and after the period conducting the research. Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind.



...Houston, we have a problem...



... is «data» still a dirty word in the SSH?

fact that data in the humanities are also an effect of operationalisation and interpretive processes.



Are artistic/humanities disciplines/methodologies only considered valid if we call them 'data'?
In other words, can we not accept different kinds of research methodologies as valid "on their own terms", rather than on those imported from (or imposed by) other fields? #munin2018

In the humanities, we all use research data, although we may not be aware of it. It is like in the case of Monsieur Jourdain, the title character of Molière's *Le Bourgeois gentilhomme*, who learnt, to his great satisfaction, that unwittingly he had been speaking prose all his life. With research data in the humanities it is exactly the same: you are using it, even if you don't know it, and once you realise it, it will affect your research workflow forever.

of Social Anthropologists. One reason for the uneven adoption of the term could be that in the humanities, "[w]e resist the blanket term 'data' for the very good reason that we have more and precise terminology (e.g. primary sources, secondary sources, theoretical documents, bibliographies, critical editions, annotations, notes, etc.) available to us to describe and make transparent our research processes" (Edmond & Tóth-Czifra, 2018:1). The resistance to 'data' in the humanities, as an oversimplifying abstraction of complex phenomena, was voiced by many critics, most notably by Johanna Drucker (2011), who opposed the objectifying term 'data' (something given) and proposed to use 'capta' (something captured, taken) instead. This

What is data?

DATA ARE NEVER «RAW» IN
THE HUMANITIES

DATA ARE ALWAYS
EXPRESSION OF A METHOD

THERE IS ALWAYS A CHOICE
[METHODOLOGICAL,
EPISTEMOLOGICAL,
POLITICAL...]

THERE IS ALWAYS
INTERPRETATION,
SUBJECTIVITY
[DATA ARE NOT GENERATED
BY A MACHINE]

DATA ARE DISCUSSED,
NOT TAKEN FOR GRANTED

ALSO BOOKS ARE DATA FOR
FURTHER
RESEARCH/INTERPRETATION

DATA=ANYTHING YOU CAN
FORMALIZE THROUGH A
LANGUAGE

BETTER «RECORD»
THAN DATA

DATA= DOCUMENTS
- WEAK (MERE
REGISTRATION)
- STRONG (HUMAN
INTERVENTION)

DATA ARE A PROCESS,
DYNAMIC AND
DIACHRONIC

Data, the pillars



CO-OPERAS

DATA ARE SO DIVERSE
(archaeological samples, pictures,
manuscripts, lexicon, corpora, ...)
THAT IT'S IMPOSSIBLE TO CONVERGE
ON A COMMON DEFINITION



CO-OPERAS

NEXT STEPS: TO DEFINE DATA, GO AS
DEEP AS POSSIBLE INTO RESEARCH
PRACTICES AND KNOWLEDGE
CREATION WORKFLOWS



We could then define data in the humanities broadly as all materials and assets scholars collect, generate and use during all stages of the research cycle. In this report we focus on digital assets.



RECOMMENDATIONS

- » Think of all your research assets as research data that could be potentially reused by other scholars. Consider how useful it would be for your own work if others shared their data.
- » Familiarise yourself with the FAIR Data Principles before you start collecting data and building corpora e.g. FORCE11: the FAIR Data Principles, GO-FAIR: FAIR Data Principles and discuss with colleagues and experts to build a better understanding.
- » Digitally document all your research and data collection work -- at the beginning of a project it is difficult to judge which information of the research process will be important and valuable later on.
- » Use well-established tools to facilitate your research work, as many of them allow data sharing e.g. MIT Libraries Digital Humanities: Tools and Resource Recommendations.
- » Browse humanities datasets and consider whether your own assets could be published in a similar fashion (e.g. Humanities Commons, UK Data Archive, ARCHE re3data.org filtered for humanities).
- » When you start producing data, keep this maxim of Open Science in mind: data should be 'as open as possible and as closed as necessary'.



[caveat]

New analytical methods and tools are being developed to exploit this abundance of data in all fields of research, leading to many questions and challenges. For instance, analysing large corpora requires the use of automated tools and methods that may not be commonly employed in humanities research methods. While tools can be used to simplify data processing, they do not always allow the fine-grained analyses required by the methodologies and theoretical frameworks employed in the humanities. Moreover, visualisation of massive datasets may highlight important, large-scale trends, but it tends to transform vast corpora of complex data into a synthetic and necessarily reduced representation of information, which can lead to the criticism that complex realities have been oversimplified in the process. It does not mean, however, that we need to be reluctant to try

new approaches or choose between the two. For instance, in textual studies many would argue that distant reading or macroanalysis (i.e. computational approaches) needs to be supplemented by close reading (i.e. informed philological interpretation of particular texts or excerpts) in order to fully appreciate the results of a massive data mining. Hence the synthesis of both approaches yields innovative results.

RISK OF
OVERSEMPLIFICATION





CO-OPERAS

Preliminary issues

IN WHICH STEP AND HOW SHOULD THE FAIR PRINCIPLES BE APPLIED?



Awareness of the FAIR principles and willingness to adopt them is not sufficient to transform data practices in any discipline. The paradigm shift requires effort, and this effort, which impacts on many roles in the research and higher education sectors, requires incentives, support, and recognition for adoption to be successful.

METADATA IN WHICH LANGUAGE? NATIONAL, SAME AS DATA, ENGLISH?

IT'S TIME CONSUMING, AND THERE IS NO INCENTIVE OR REWARD

LACK OF DATA AND METADATA SKILLS AMONG RESEARCHERS

WE NEED TO PRESERVE THE SPECIFICITY OF HOW WE DO RESEARCH IN THE HUMANITIES

ABSOLUTE NEED OF A REGISTRY OF EXISTING TOOLS

SERVICES AND TOOLS NEED TO BE SUSTAINABLE

ARE WE, THE RESEARCHERS, WHO USE THESE RESOURCES WITHOUT ASKING HOW DATA WERE CREATED OR DESCRIBED, PART OF THE PROBLEM?

To be FINDABLE

METADATA ARE CRUCIAL,
AND THEY CAN NOT BE
SEPARATED FROM DATA

MAINTAIN THE RICHNESS OF
METADATA – IF YOU LOSE THE
CONTEXT IN THE SSH, YOU
LOSE EVERYTHING

ALMOST COMPLETE LACK
OF METADATA
STANDARDS

IF WE USE LARGE SCALE
REPOSITORIES, THEY DON'T
PROVIDE THE EXPECTED
GRANULARITY
(DESCRIPTION/DISCOVERABILITY)

LACK OF UNIQUE
IDENTIFIERS

UNEVEN METADATA
QUALITY

LACK OF
INDEXES/THESAURI

[why is it so difficult?]

The Scholarly Data Continuum

The previous sections highlighted that, in contrast to the hard sciences, the initial data in the arts and humanities is collected²⁴ rather than generated,²⁵ and thus the digitisation of cultural heritage is an indispensable base for research in these disciplines. However, considering the highly intertwined

A Cultural Knowledge Iceberg, Submerged in an Analogue World

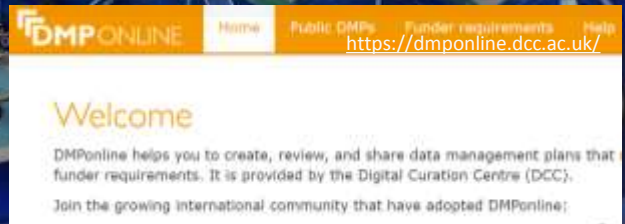
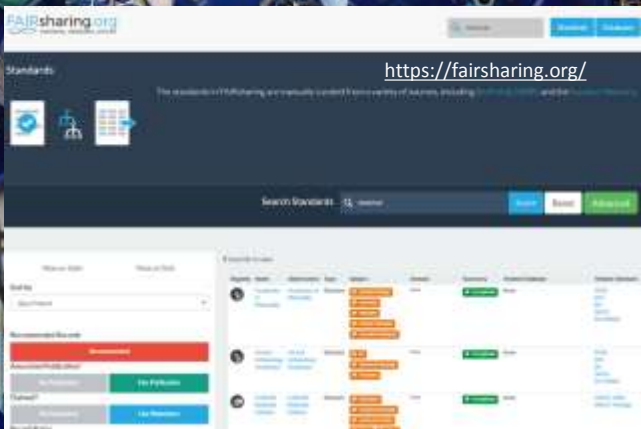
There is a fundamental difference between the epistemic cultures of STEM (science, technology, engineering, and mathematics) and those of the arts and humanities: namely, that in the arts and humanities the wide range of scholarly information artefacts, works of art, written documents of all sorts, recordings, annotations etc. — all of which can be broadly referred to as research data (in the sense used by Margaret E. Henderson)⁷ — are not the autonomous products of research projects, but rather are deeply embedded in the cultural memory of Europe as well as the cultural and social practices of the institutions that preserve, curate, and (co)produce them. These institutions, commonly referred to as cultural heritage or GLAM (galleries, libraries, archives, museums) institutions — ranging from national libraries and archives down to small village museums or administrations — are typically not part of the institutional landscape of academia. Despite this, the digital research ecosystem poses many challenges connected to the exploration and exploitation of the material and collections they hold; we do not need to get very far into the FAIR acronym to recognise these challenges.

- DATA ARE COLLECTED RATHER THAN GENERATED
- DATA ARE NOT AUTONOMOUS PRODUCTS BUT THEY ARE EMBEDDED IN THE EU CULTURAL MEMORY

A speciality of data management in arts and humanities, therefore, is that it is highly dependent on external data providers, that is, the cultural heritage institutions.²⁷ As was also touched on in the

accessibility of complex knowledge structures. As a result of the separation of data from its context of creation (i.e. from the institution, its curators, and its wider provenance), collection descriptions that are part of the standardised and aggregated metadata remain the only reference points for the long history of records.

Findable - Me



RECOMMENDATIONS



- » If applicable, determine if the body funding your research has particular requirements for a DMP, or offers a template for framing your plan. If there is no required template, choose an existing appropriate one (e.g. via DMPOnline).
- » Devise a DMP prior to collecting data. Define and plan for your data: all research projects deal with data. If your project includes the analysis of text corpora, for example, then the corpora themselves are data, and you should make sure they are clearly described, documented, and managed according to the FAIR principles so your research is reusable by others.
- » Plan documentation of metadata: In order for your data to be comprehensible in the future and/or reusable by others, they will need descriptive metadata created according to a common schema to understand the context/purpose of the research. The richer the metadata, the more intelligible and useful the dataset (see section on Metadata).
- » Use standardised terminology to increase interoperability. Consider employing vocabularies or ontologies that follow FAIR principles to increase interoperability and findability (e.g. see FAIRsharing.org).
- » Consider the right questions to be answered in your DMP that can account for discipline-specific requirements. The DMP templates suggested by funders are quite high level and provide generic guidance for file naming or versioning conventions, database structuring, and can be a good start. Tools like the dmponline.dcc.ac.uk provide discipline specific examples that can be of further reference.
- » DMP as living documents: Update your data management plan regularly in order to take into account any potential relevant changes such as using new data types and/or models, technology, new institutional data management policies, reassessing legal aspects or licences for legal compliance etc.

Next steps

EXPLORE A MINIMAL METADATA SET, EXPANDABLE [TO BE ADDED TO FAIRSHARING]



CO-OPERAS

CUSTOMIZE DATA STEWARDSHIP WIZARD FOR SSH AND SPECIFIC DATA [OUTPUT] MANAGEMENT PLANS



Chapters

- I. Administrative details 5
- II. Re-using data 1
- III. Creating and collecting data 8
- IV. Processing data 7
- V. Interpreting data 8
- VI. Preserving data 10
- VII. Giving access to data 3

II. Re-using data

Before you decide to embark on any new study, it is nowadays good practice to check all options to re-use existing available data, either collected or generated by yourself in an earlier project, or data from others (Barend Mons calls this "Other PEople's Data And Services" or OPEDAS). This can include reusable data that have been created for an earlier study, and also so-called "reference data" which is used by many projects.

It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money is bringing with it the responsibility to treat those data well and (if potentially useful) make them available for re-use by others. And the circle is only complete if such data is actually re-used.

1 Is there any pre-existing data?

Are there any data sets available in the world that are relevant to your planned research?

Data Stewardship for Open Science: [atq](#)

a. No

b. Yes

Start planning

Install

Contact us

- DISCIPLINARY WORKSHOPS
- SHARE BEST PRACTICES
- WORKSHOPS ON FAIRIFICATION OF REAL, EXISTING DATABASES

To be ACCESSIBLE



CO-OPERAS

LACK OF A
DISCIPLINARY
REPOSITORY
(WITH ADEQUATE
GRANULARITY OF
METADATA)

MULTILINGUALISM

NEED OF SPECIFIC
SOFTWARE TO
READ/TEXT AS
IMAGES NOT
SEARCHABLE



LACK OF SINGLE POINT OF
ACCESS TO TEXT, DATA

PHONETIC
TRANSCRIPTIONS

NEVER STOP MAKING IT CLEAR...
ACCESSIBLE DOES NOT EQUATE TO OPEN
ACCESSIBLE =
WHERE TO FIND THE DATA
UNDER WHAT ACCESS CONDITIONS

Accessible

TRUSTED REPOSITORIES IMPORTANCE OF CONTEXT

For data to be managed over the long term, and made accessible in a continuous and sustained way, it should be deposited in a location that ensures trusted, ongoing stewardship of the data. Researchers depositing their data and those accessing it for reuse should be assured that data sets are authentic, retrievable, annotated sufficiently to understand the **context** of their creation, and assigned licence information that clarifies the conditions of reuse.

DIGITAL PRESERVATION

However, storage is not the same as preservation, because digital data are fragile and subject to corruption and degradation over time. File formats or the software and hardware required to access them may also become obsolete. Data published on websites can become inaccessible when links break, pages are moved, or the website disappears. Over time, technology, human actions (or inaction) and environmental factors challenge the integrity of data, so simply 'backing up' that data is not sufficient: it must be preserved. Digital preservation is not a single action, but a process that is designed to **ensure digital data are continuously accessible** into the future, through all of the changes that time and technology can inflict.



RECOMMENDATIONS

ALLEA report, p. 28



- » To ensure the best possible stewardship of your data, choose to deposit it in a digital repository that is certified by a recognised standard such as the CoreTrustSeal. The Registry of Research Data Repositories (re3data) provides a good starting point, noting disciplines, standards, content types, certification status and more. FAIRsharing (manually curated information on standards, databases, policies and collections) allows you to search databases by subject, and includes entries tagged 'Humanities and Social Sciences'.
- » Use disciplinary repositories where they exist, as they are more likely to be developed around domain expertise, disciplinary practices and community-based standards, which will promote the findability, accessibility, interoperability and ultimately the reuse and value of your data. The level of curation available in a repository is key to data quality and reusability.
- » Datasets should be assigned persistent identifiers (PID). Most repositories that are designed for long-term preservation will automatically assign or 'mint' persistent identifiers for your datasets, so choosing a quality repository will automate this step. Consider as well signing up for ORCID, a free service that assigns persistent identifiers to individuals/authors.
- » To facilitate findability of all research outputs, bidirectional links should be created between publications related outputs, such as data (using PIDs).
- » Include the richest metadata possible with your deposited data so that others can find it, understand the parameters under which it was created, and understand the conditions under which they can access and/or reuse it. See recommendations in this report in the sections on Licences and Metadata for more information.

BIDIRECTIONAL LINKS

PIDs also facilitate citation, and for increased findability, links should be created between publications and their associated datasets (bidirectional linking). These links are often created

To be INTEROPERABLE

METADATA ARE SO TAILORED
ON A SPECIFIC PROJECT THAT
INTEROPERABILITY IS ALMOST IMPOSSIBLE

FORMAT CONVERSION
(POSSIBLE LOST OF
FUNCTIONALITIES OR
INFORMATION)

COMPLEXITY OF CORPORA
(THEY HAVE MANY LAYERS)

ABSOLUTE NEED OF
ONTOLOGIES
(PHILOSOPHICAL JOB, NOT I.T.)

VOCABULARIES AND
SEMANTIC ARE CRUCIAL

SUGGESTION FOR
FUTURE HACKATHONS!

Interoperability



RECOMMENDATIONS



» Prefer human and machine-readable systems: coding of data models and of the actual data that is both human and machine-readable in a unified way provides better sustainability and long-term accessibility than machine-readable only code (binary codes), that may use different formats for data model description and the actual data. For both, hierarchical data models and graph-based data, various serialisations (file formats) are available that fulfil this condition (XML, TEI/XML, Turtle, N3, RDF/XML), whereas SQL based technologies need bigger efforts.

» Normalise as much as possible: to avoid redundant information, the content of databases should be normalised as far as possible, using for example authority files like VIAF and identifiers like DOI, ARK, ISNI, GND and the like. To foster the exchange of data, standardised vocabularies and ontologies are needed as well, but an overall ontology for the humanities has not yet been established. The ontology CIDOC-CRM and especially some extensions are well on their way to become a reference model for cultural heritage information.

» Data models follow the data management plan (DMP): when establishing a data model, researchers should keep the whole lifecycle of their data in mind, as it should be outlined in a DMP. Therefore, an extensive documentation of the data model, its software and tools are highly relevant and facilitates the transfer of data in a secure and trusted repository in order to keep them accessible. The same is true here: the more you use open standards for your data model, the easier this task becomes.

» Data models go FAIR: the FAIR Guiding Principles, correctly applied, ensure data are findable, accessible, interoperable and reusable. Data modelling should take this into account by using formal, easily accessible languages for knowledge representation, providing persistent identifiers, open standards, well documented Application Programming Interfaces (API), generic user interfaces and rich metadata. The FAIRification process developed by the GO FAIR initiative offers a system on how to shape the data modelling.

» Use open standards, and whenever possible, standardised technologies and procedures should be used. The World Wide Web Consortium W3C maintains several standards relevant for data models like XML and RDF. Within XML the Text or Music Encoding Initiative TEI/MEI or specific expressions of them have become standards for text or music editions. The query language SPARQL and the representation tool for linked data JSON-LD are common standards for RDF (refers to FAIR principle I1).



To be Reusable

DATA IN THE HUMANITIES OFTEN
HAVE A LEGAL UNCERTAINTY



Legal Problems that Are Not Solely Legal Problems

The biggest obstacle in the productive reuse of digitised cultural heritage resources — from which many others derive — is the legal and ethical restrictions in which the usage conditions of cultural heritage sources are embedded. Determining the ownership status of research that is based on such material poses challenges in many cases. This is because the ownership status of research is, on some level, shared between the researcher who carries out the scientific analysis on the source materials, the institution that hosts and curates this material, and the people and cultures who give rise to the objects in question (e.g., photographers, and also the subjects of the photographs). Establishing precise conditions for reuse on the basis of such a complex web of claims is, therefore, not an easy task.¹²

In addition to this complexity, provenance trails (i.e. a documented ownership and curation history of an artefact) are often embedded in historical practices, in particular in eras or contexts when the legal-ethical framework that defines present-day data exchange was either non-existent or irrelevant. Obviously, those handling these data

COPYRIGHT IS STILL AN
ISSUE

NEED OF LEGAL ADVICE ON
LICENSES

BUT, IN THE END, REUSE IS
THE FINAL AIM OF ALL THIS
PROCESS!!!

AND REUSE CAN ALSO BE A
MEASURE OF THE SUCCESS
OF A RESOURCE

Reusable / rights

Data generated or collected in the e-humanities may potentially be subject to copyright in whole or in part. In addition, particular challenges may arise when some of the items within a data set themselves are subject to third party rights.

- Which national legislation applies to other researchers' work I use in my project?
- Do I have the right to collect, preserve and provide access to the data of my project?
- Is there sensitive information that could connect to some privacy issues?
- Are there risks of exposing the identity of human participants in my study?
- Am I allowed to digitally reproduce material and (re-)publish it in a digital reproduction?



RECOMMENDATIONS

ALLEA Report, p. 24-25



- » Clarify all legal issues at the beginning of your research project and include the findings of this process in the data management plan.
- » Use checklists adequate to your research topic/discipline.
- » Check the resources indicated by DARIAH, CLARIN. (see further reading).
- » In the case of personal data ensure that only relevant people can access the data and that these are clearly identified (see GDPR).
- » Ask for consent to share anonymised data and establish transparent and well-documented anonymisation routines that consider not just direct identifiers, but also how a combination of indirect identifiers could reveal identities. (See for example the guide on informed consent in the CESSDA data management expert guide).
- » Avoid collection of (sensitive and non-sensitive) personal data when possible.

Reusable / licenses



RECOMMENDATIONS

ALLEA Report p. 26

Researchers are “prosumers” who produce and consume information and knowledge of other researchers. This section focuses on their role of producing knowledge and on ways to foster its diffusion by clear legal boundaries. In the humanities, texts are quite often closely intertwined with underlying data, which form an indispensable part of digital publications. Traditional conceptions of copyright like “All Rights Reserved” raise obvious problems for data sharing in the context of publications. In general: if machine readable data is to be shared, the recipient, in order to use the data effectively, will most likely need to make a local copy for analysis, or for merging with other data sets, or to extract some subset of the data. For this reason, our recommendation is to avoid applying any legal restrictions that do not embrace the principle of openness. The Reusability FAIR

» Proper entitlement: first of all, identify who owns the data, i.e. whether you are entitled to license your work. You may only attribute a licence to a work of which you are the copyright holder. If there are co-authors, you have to agree with them on the licence. Furthermore, you are not allowed to license the works of the public domain. You should also be aware of whether there are any licensing requirements from the funding organisation or the data repository.

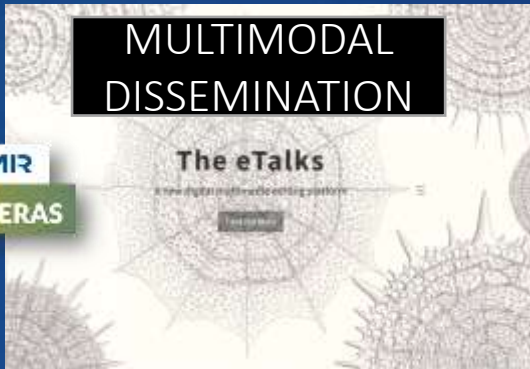
» Determine the necessary and sufficient level of access restrictions. Some data cannot be shared openly but can still be shared under certain restrictions while at the same time protecting the data. See for instance the [CLARIN licensing framework](#) for language data or the [CESSDA access categories](#) for qualitative and quantitative data (interviews, survey data etc).

» Use free and standardised licences: In order to benefit from the possibility of sharing data since the digital turn and to foster Open Science, use a licence as free as possible. The Open Knowledge Foundation and the [Open Access Scholarly Publishers Association](#) only

» For editors of journals and repositories managers: Avoid applying more restrictive licences like NC (non-commercial) or ND (no derivatives) just to be ‘on the safe side’. NC can produce unintendedly limiting side-effects to potential re-users, as it is not quite clear whether the setting of a re-used work has commercial aspects or not. ND originates from the creative sector and is thought of as an instrument to protect the integrity of a work of art, such as a music composition. Many humanities scholars also want to protect their works from misuse and therefore are in favour of a ND licence. However, the risk of misuse through derivatives in the humanities is often quite low, so one has to balance this potential risk against the potentially unintended constraints imposed by ND, such as restrictions against reuse of publications in text and data mining procedures. Keep in mind that anybody deliberately deriving original content and thoughts by other scholars with misleading intention violates ethical scientific behaviour, whether a work is put under and ND licence or not.

As open as possible, as closed as necessary

MULTIMODAL DISSEMINATION



eTalks



RECOMMENDATIONS

ALLEA Report, p. 32

» Humanities scholars are encouraged to take advantage of the frameworks, networks and resources that facilitate the discoverability and wider reuse of research:

- Domain registries, portals, harvesters, e.g. [Re3data](#) and [FAIRsharing.org](#)
- Platforms e.g. [Europeana](#), [AGATE](#)
- Researcher profiles e.g. [ORCID](#)

» Share online your data and all supporting materials such as presentations, posters, blogs, data papers etc. and consider using social media for wider outreach, cite using persistent identifiers.

» Consider publishing a data paper either as a preprint or via a dedicated journal for data papers. An emerging practice supporting the FAIR principles, publishing data papers about data sets increases findability as well as reuse, as these provide the key information about specific datasets. e.g. [Journal of Open Humanities Data](#), [Research Data Journal for the Humanities and Social Sciences](#).

» Talk about your research outside academia, consider diverse audiences, such as journalists, policy makers, private companies or citizen scientists as Open Science is ultimately promoting the involvement of a wider audience in scientific research.

» Consider non-traditional channels and formats to present your data: infographics or interactive data visualisations, online exhibition or digital tours, websites or apps, executive summary/lay summary, also consider a wider use of national languages.

» Promote/prepare your datasets for use in class (schools or HEI) or for Hackathons (e.g. [Coding Da Vinci](#)).

» As an institution, actively also showcase and provide institutional channels that researchers can leverage, and reward data dissemination.

» Encourage and support pedagogic approaches which include student production and curation of open research data, and use of existing open datasets as open educational resources (OER).

as a starting point. Active dissemination around data, once the data have been made FAIR, needs to become a key research data management best practice.

The rapid uptake of the FAIR data principles as part of a wider movement towards Open Science is changing how scientists and scholars collect, curate, preserve and share their research data. In particular the principle of “as open as possible, as closed as necessary” is aimed at guiding researchers in their efforts to strike a balance between sharing data and the need to account for issues around sensitive data/legal aspects. Overall this shift has also brought a focus on maximising data use and potential not only for future research but also in other areas (e.g. private sector) and for other categories of potential users (e.g. citizen scientists).



. thank you and... call us!

elena.giglia@unito.it

T.Biro@ria.ie

www.operas.eu

www.allea.org