DARIAH-EU

RESEARCH DATA MANAGEMENT FOR ARTS AND HUMANITIES

INTEGRATING VOICES OF THE COMMUNITY

Research Data Management Working Group

RESEARCH DATA MANAGEMENT FOR ARTS AND HUMANITIES

INTEGRATING VOICES OF THE COMMUNITY

Review:

The publication underwent an open peer review process whereby there was a period dedicated to the Working Group members reviewing and commenting on each other's subchapters and discussing their work. All 14 authors participated in the process.

Graphic and cover design, typesetting:

Agnieszka Zalotyńska

Proofreading:

Laurence Taylor

Digital Object Identifier (DOI):
10.5281/zenodo.8059626

Further Resources:

This work is also available as an interactive online publication here:
gitlab-ce.rrz.uni-hamburg.de/uahh-digitale-dienste/rdm-for-arts-and-humanities.
Bibliography can be downloaded here: www.zotero.org/groups/2427138/collections/5H29SAX6.

Suggested Library of Congress Subject Headings (LCSH)

Digital humanities (id.loc.gov/authorities/subjects/sh2008122106)
Research--Data processing (id.loc.gov/authorities/subjects/sh85113022)

Funding:

The publication and the writing sprint were made possible by the third Working Groups (WG) Funding Scheme Call for the years 2021-2023 and the resulting grant administered by the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN), with the project title: *Research Data Management for Arts and Humanities: Integrating Voices of the Community*.

Contact:

wg-rdm@dariah.eu

August 2023

Working Group Chairs and publication editors:

Erzsébet Tóth-Czifra
Marta Błaszczyńska
Francesco Gelati

Authors[1]:

Erzsébet Tóth-Czifra, ORCID: 0000-0002-5350-067X
DARIAH-EU (Berlin) / CoARA (Strasbourg)

Marta Błaszczyńska, ORCID: 0000-0002-2377-4565
Instytut Badań Literackich Polskiej Akademii Nauk (Warsaw), ROR: 02q0p6x28

Francesco Gelati, ORCID: 0000-0002-6066-1308
Universität Hamburg, ROR: 00g30e956

Femmy Admiraal, ORCID: 0000-0002-3609-7073
Universiteit Leiden, ROR: 027bh9e22

Mirjam Blümm, ORCID: 0000-0003-3665-7031
Technische Hochschule Köln (Cologne), ROR: 014nnvj65

Erik Buelinckx, ORCID: 0000-0003-1831-158X
Koninklijk Instituut voor het Kunstpatrimonium - Institut Royal du Patrimoine Artistique (Brussels), ROR: 01phtp995

Vera Chiquet, ORCID: 0000-0001-5925-2956
Universität Basel, ROR: 02s6k3f65

Rita Gautschy, ORCID: 0000-0002-5470-8720
Universität Basel / Swiss National Data and Service Center for the Humanities (Allschwil), ROR: 047f01g80

Peter Gietz, ORCID: 0000-0002-8310-2015
DAASI International (Tübingen)

Péter Király, ORCID: 0000-0002-8749-4597
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, ROR: 00cd95c65

Maria Vivas-Romero, ORCID: 0000-0003-2196-6923
Universiteit Maastricht, ROR: 02jz4aj89

Walter Scholger, ORCID: 0000-0002-9256-0958
Karl-Franzens-Universität Graz, ROR: 01faaaf77

Bartłomiej Szleszyński, ORCID: 0000-0002-7758-1662
Instytut Badań Literackich Polskiej Akademii Nauk (Warsaw)

Ulrike Wuttke, ORCID: 0000-0002-8217-4025
Fachhochschule Potsdam, ROR: 012m9bp23

[1] About ORCID, see: orcid.org. About ROR, see: ror.org.

Table of Contents

INTRODUCTION

Erzsébet Tóth-Czifra

'At that time, I was a mere historian [...] So at that time, I was happy I just could do my research and didn't have to mind the standardisation and research data management and so on; aspects that became such a thing later on. So it just went into the process of our institution using only the prosopographical data. Now I think we can do more with it'.

[excerpt from an interview on sharing historical data work]

The excerpt above might have some resonance with many of us, as it gives an honest account of the changing expectations and the eventually changing practices of how research is designed, carried out, disseminated, and documented. Working with sharing in mind, uncovering research processes, and making underlying resources available so they become truly accessible building blocks which others can (re)use, brings about a massive cultural change, leaving none of the disciplines, or their broader homes and research institutions, untouched as new priorities and roles emerge. Addressing the growing need to enable technological and cultural solutions for sustainable data sharing as well as to comply with growing European-level and national data policies, research data management has emerged as a new field of expertise to be explored and established across the whole range of disciplines – and now there is time to reflect on the first period of our familiarisation with FAIR principles, data management plans, vocabulary standards, archive-friendly formats, data repositories, and a range of other elements of new research, research support, and advocacy realities, all of which we will eventually live by.

About the DARIAH Research Data Management Working Group and this publication

The idea behind the DARIAH Research Data Management Working Group is to tackle the challenges associated with the implementation of new FAIR and open data sharing mandates, and offer a unique space for collaboration among representatives of all major arts and humanities disciplines, cultural heritage professionals, and data management experts. More specifically, we are building a knowledge hub for new professionals around data management support (data managers, data stewards, open science officers, subject librarians etc.) from across DARIAH's national hubs to exchange across the discipline-specific dimension.

During our meetings, the conflict between the need for and emergence of new data support roles, and the lack of any established domain-specific curriculum to train them, or, in some cases, even a lack of established good practices, became a recurrent topic

which we aimed to address. Instead of embarking on the giant endeavour of curating an exhaustive and authoritative textbook for research data support specialists who work in the arts and humanities field, our idea is to give a snapshot of our own activities in the field and highlight the valuable work of others. Accordingly, the present publication can be read as the written form of a roundtable (or town hall) discussion where experts from Austria, Belgium, Germany, Switzerland, the Netherlands, Poland, and Spain have come and sat together to report and discuss past or ongoing work, share their fields of interests and provide honest and critical reflections, reveal how their institutions have developed capacities for data support, share their own stories of becoming data support professionals in the domain, and, most importantly, explore together what solutions, tools, practices, and other resources could be used and could be generalised across borders and disciplines.

Chapter one gives an overview of the European and national policy environment which has given rise to research data management and sharing mandates, as well as the institutional support structures around them. In chapter two, which is dedicated to implementation and everyday practice, the authors of this publication share how their institutions have developed capacities to accommodate data support professions, and also share their own career paths leading to such roles. After the first two chapters have set the stage and recounted the authors' reflections on these new roles, the rest of the publication highlights and discusses some of the key domain-specificities of research data management in the Arts and Humanities. Chapter 3.1 reflects on the implications of the lack of consensus around the notion of data within the Arts and Humanities domain through a case study of digital critical editions. Chapter 3.2 addresses the challenges around the, essentially, multilingual character of arts and humanities data, with special focus on multilingual vocabularies and thesauri. Chapter 3.3 provides support for research scenarios where open data sharing is either impossible or is difficult due to legal and ethical limitations, and navigates the complexities of intellectual property and the application of regulatory frameworks, including restrictions on text and data mining, and authentication and authorisation in an open world. Clearly, the discourse on data sharing cannot be complete without discussing the current limitations within research assessment and rewards criteria, nor highlighting initiatives which aim to incentivise and reward data sharing in the working/professional contexts of the Working Group's members. A discussion on rewards can be found in Chapter 3.4. Chapter 3.5 addresses one of the most widely shared data management challenges within the domain and brings together use cases concerning successful collaborations between cultural heritage institutions and arts and humanities research teams. Finally, Chapter 3.6 showcases good practices in long-term archiving.

This publication is the result of the Working Group's first writing sprint, held between the 23 and 24 June 2022 at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN), in the Staszic Palace. The event was made possible by the third Working Groups' (WG) Funding Scheme Call for the years 2021–2023 and the resulting grant, which was administered by the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN).

There were two key considerations which led to making our work available in a GitLab book format. The first one has to do with practising what we preach in terms of data management and making our work available in a structured format (markdown). Second, this type of format is dynamic and flexible enough to enable the creation of an interactive book format and to accommodate comments, additions, and possible corrections from the broader community who are working in such a dynamically evolving field. Thus, we invite our readers to join in the discussion and make their voices heard!

In addition to the dynamic GitLab version, a static PDF version of this publication, serving as a version of record, is available on Zenodo.

1. A BRIEF OVERVIEW OF THE RESEARCH DATA MANAGEMENT POLICY LANDSCAPE

Marta Błaszczyńska, Mirjam Blümm, Erik Buelinckx, Vera Chiquet, Rita Gautschy, Francesco Gelati, Erzsébet Tóth-Czifra, Ulrike Wuttke

In this chapter, we provide our collective understanding of how research data management (RDM) and its evolving policy landscape became 'such a thing', what the major entailments are for research institutions in general and arts and humanities research in particular, how it sparks new ways of collaboration between researchers and cultural heritage institutions, and how it brings about the new data support roles which many of our working group members fulfil.

Our reflections come from different roles and from a diverse range of institutional and national contexts and perspectives. The plural noun 'voices' in the title of our publication, *Integrating Voices of the Community*, stands to highlight – even celebrate – this polyphony; while 'community', in the singular, showcases the fact that at the intersection of the arts and humanities research domain (an umbrella term in its own right) and emerging research data support roles (under the titles of data stewards, open science officers, subject librarians and many others, across Europe), a new community is coming to life with shared professional interests, challenges, and a strong wish to foster exchange and consolidate discourses and good practices at the European level. The DARIAH Research Data Management Working Group aims to provide a central hub for this community.

1.1. The evolution of FAIR principles in Europe

In just the couple of years following their inception in 2014, the FAIR principles (Wilkinson et al. 2016) became key components in research data mandates across Europe. Despite their top-down character, that is, the broad embrace and support of FAIR by governments, policy-makers, governing bodies, and funding bodies, the rapid uptake of FAIR principles by these bodies can still be interpreted as a response to the increasingly digital, scientific, and research economies, bringing with them a strong and special need for guidelines which enable and incentivise sustainable, connected, easily accessible, and cost-effective models of scholarly data curation, and which can also be extended to data types which naturally fall outside the scope of open data mandates, such as sensitive data, ethically fragile content, or copyrighted materials.

Figure 1. *9412 citations of the FAIR principles as of 01.04.2023.*

Although early critics of the FAIR principles (such as Skatz 2017; Mons et al. 2017; Mons et al. 2020) emphasised the highly schematic and interpretative nature of the principles (the most concrete formulation breaks down the principles into 15 sub-principles, see www.go-fair.org/fair-principles/) as well as their foreignness in relation to these principles in actual research realities, FAIR, as a policy instrument, has undeniably given rise to changes in research cultures over the past years.

Most notably, it has fostered discussions about standards, data sharing, and licensing practices in disciplines which, traditionally, are not considered data-driven or standardisation-heavy – such as many of the disciplines belonging to the arts and humanities domain (Harrower et al. 2020; Hettne et al. 2018). Second – and this is partially in response to those critics who point out the costs as well as the crucial infrastructure and capacity preconditions of implementing FAIR data mandates – ever since the publication of the *Turning FAIR Data Intro Reality* European Commission report, we see new research data-support structures emerging across Europe. Some of them come with limited time frames, such as funding resources allocated to FAIR data and software implementation projects both at the European and national levels.

Another group of these structures has given rise to long-term infrastructure commitments (NDFI in Germany, see discussion later in this chapter, and most notably the European Open Science Cloud (EOSC)) and, around them, new professional data-support roles, bridging the gap between research policy and research infrastructures and research realities. Although the *Turning FAIR Data Intro Reality* report refers to the latter as data stewards, in reality, they are operating under very diverse names (open science officers, FAIR data officers, subject librarians etc.) and with even more diverse capacities across Europe. In Chapter 2, our authors – most of them fulfilling these new research data support roles within in the arts and humanities domain – share insights and personal testimonies about how their institutions have developed capacity for these new roles, how they found their way into them, what their daily challenges are, how their first-generation experiences might shape the future of these professions, and, most importantly, how they bring research data management to disciplinary realities. The following section adds a little context to this area.

1.1.1. FAIR principles and epistemic cultures in the Arts and Humanities

As mentioned above, the rapid uptake of FAIR principles in policies has brought about discussions concerning data management, licences, PIDs, and data publication to a range of disciplines, arts and humanities disciplines among them, without these discussions being burdened by established traditions of thinking in terms of data, data sharing, or standards. This, by far, does not mean that early attempts at implementation did not spark caution, or even resistance in some cases, in data- and standards--savvy and less digitally oriented disciplines.

The reasons for this are manifold, starting with the term 'data', which is still a problematic umbrella term, alien to many of the epistemic traditions in the Arts and Humanities (see, e.g, Edmond et al. 2022, and also the discussion in Chapter 3.1.); that data is collected and curated rather than generated (see Tóth-Czifra 2020), and, therefore, alternative approaches like the CARE principles – emphasising the values of provenance, integrity, ethical complexities, and power relations inherent in digital data curation – seem to be more appropriate in such scenarios (O'Donnell 2022).

Further, in actual research and research support practice, inequalities in access to resources and support structures for the implementation of FAIR data mandates vary greatly along both geographical and disciplinary lines. In this respect, we see reflections, critical interventions, and assessment exercises surrounding Digital Humanities as well (Harrower et al. 2020; Tóth-Czifra 2020; Castro 2020), highlighting conceptual, financial, and infrastructural challenges, but also opportunities in realising the promises of FAIR within discipline-specific scholarly practices. Antonio Rojas Castro's recent publication, *FAIR Enough? Building DH Resources in an Unequal World*, contains a critical assessment of FAIR in the context of a text-oriented, digital humanities project called Humboldt Digital (a project aimed at the philological development and computational analysis of sources relating to Alexander von Humboldt's Cuban research within the period 1800–1830)[1], and a virtual learning environment for humanists – the Programming Historian.[2] Apart from illustrating how the 15 sub-principles of FAIR manifest themselves in these two projects, he also touches upon a range of ethical and sustainability issues, for example, how not to replicate dominant power structures while creating technologically mature, FAIR resources; how to overcome inequalities in access to infrastructure, training, and skills; to what extent funding is a prerequisite for FAIR compliance; how to overcome creator-centric approaches when describing cultural

[1] habanaberlin.hypotheses.org
[2] programminghistorian.org

resources; and who is responsible for making such decisions within a research project. Still, and contrary to the dominant impact of the natural sciences and the importance of epistemic cultures in FAIR implementation (Beyan et al. 2020), there is a growing need for scholarly communities to translate the global open data mandates into their own needs and to find their own domain- and discipline-specific solutions and paths towards responsible and sustainable research-data sharing. Using FAIR's momentum and dedicated resources to address the long-standing challenges our research domain is facing, such as improving access to cultural heritage data (see Tasovac et al. 2020), and building on standardisation efforts older than FAIR, is crucial. In light of the evolution of FAIR data policies across Europe (detailed below in the next section), through such efforts, it is not only the inclusiveness of FAIR which is at stake but also the future funding situation and well being of arts and humanities research. It is clear that we need to make it easier for our communities to comply with such policies in ways which make sense in their respective research contexts. The DARIAH RDM-WG group is committed to such translation work by building bridges between epistemic cultures and disciplinary realities.

1.2. FAIR and open data mandates in European and national funding frameworks

1.2.1. Europe

'As I see it, European success now lies in sharing as soon as possible [...]. The days of open science have arrived'.

Speech by Commissioner Carlos Moedas at the Presidency Conference Open Science, Amsterdam, 4 April 2016.

The emergence of the open science and open access paradigms came about hand in hand with the advent of the digital dimension to research. Some of the community-driven efforts and key infrastructural components such as arXiv (originating in Cornell University in 1991, and now used globally as the most widely known preprint server) and the Directory of Open Access Journals (DOAJ, founded in 2003) became essential enablers of open access and open science globally.

Having recognised the economic, societal benefits, and innovation potentials of open science, the European Commission has, since 2008, been active in implementing mandates, first for open access and then later for broader open science in European research funding programmes.

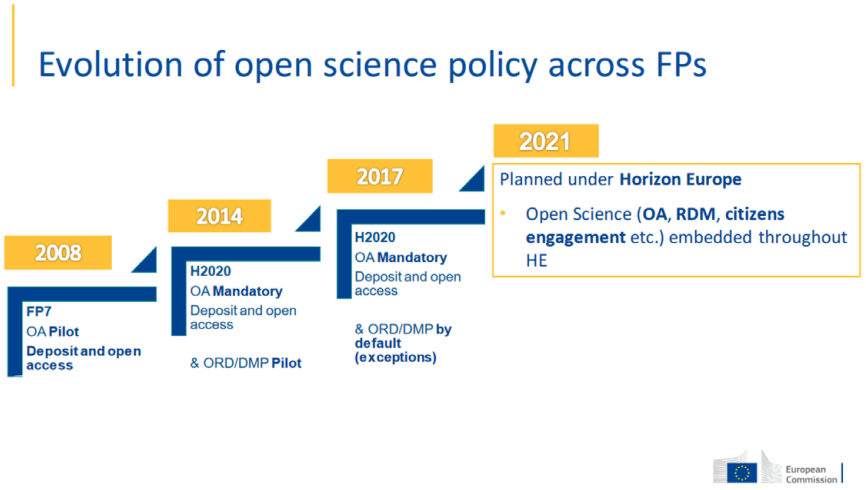


Figure 2. *Screenshot from slides presented at the Stakeholder workshop – Novelties in Horizon Europe MGA*.

The figure above sums up the evolution of the European Commission's open science policy across funding programmes, including the Horizon programme and other European grant schemes, such as ERC grants or Marie Curie grants. As can be seen, data sharing has become increasingly more salient since 2014; beginning with the advent of open research data and data management pilots which became real conditions for funding imperatives from 2017 on. The Horizon Europe framework (2021–2027) brings about a more explicit and more elaborate open and FAIR data mandate, and encompasses the broadest possible definition of research data – information should be provided about all the other scholarly objects, tools, and instruments which are needed to validate the conclusions of the publication. Research data management and responsible data sharing are now mandatory parts of Horizon Europe's projects – the open data and RDM pilots started in H2020, and FAIR data sharing are now expected to be the default components for the funding program. Further, RDM costs are eligible for reimbursement from the research grants. It is still an open question as to whether this will be restricted to the project's lifetime only or could also include long-term data management costs, or, instead, beneficiaries could be supported by central services dedicated to this purpose. The *Horizon Europe Model Grant Agreement* is complemented by the DMP guidelines and templates.

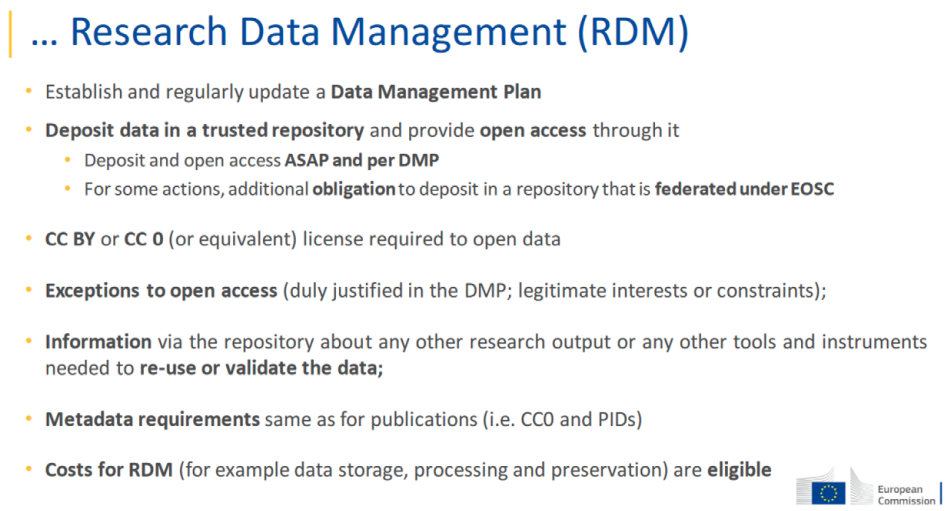


Figure 3. *Screenshot from slides presented at the Stakeholder workshop: Novelties in Horizon Europe MGA.*

To capture, collect, and consolidate good practices around data management planning, the DARIAH RDM WG curates a Zenodo library dedicated to DMPs for Arts and Humanities research and infrastructure development projects.

Looking ahead, as a likely future trajectory, we expect that connecting or integrating scholarly resources such as data and publications, services, and infrastructure components with the European Open Science Cloud (EOSC) will soon become an even stronger policy imperative than it is currently.

Another important line of policy work which will likely shape the research data landscape of the future is the broader European Data Strategy and its legislative (sub)components, the Open Data Directive, the proposed Data Act and Data Governance Act, the Digital Services Act, and the Digital Markets Act. At the time of writing this draft, there are ongoing consultations around what possible impacts each of these will have on research and research data management. As a flagship infrastructural commitment, the European Commission, in 2021, launched calls for the development of common European data spaces. These data spaces will ensure that more data becomes available for use in economies and society, while keeping the companies and individuals who generate the data in control.

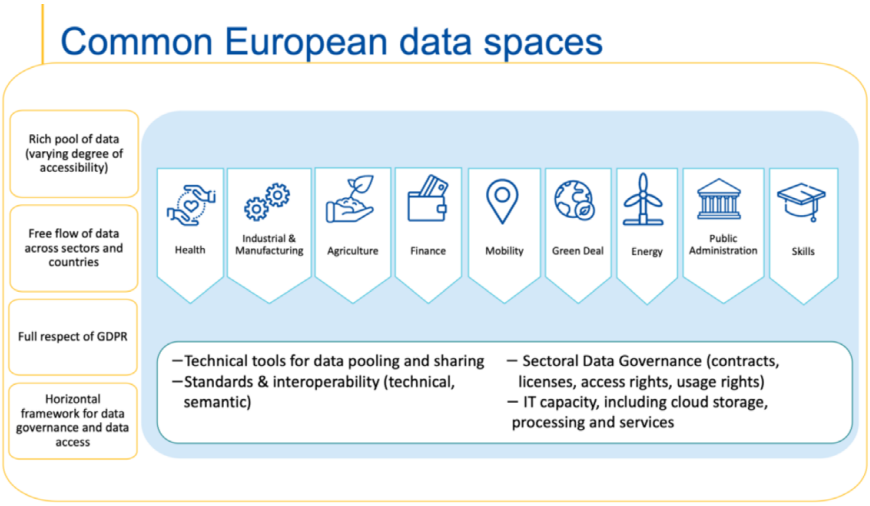


Figure 4. *Common European data spaces*[3]

The envisaged common European Data Space for Cultural Heritage is an especially relevant data infrastructure which is being developed in the context of arts and humanities data workflow, bearing in mind the essential dependence of arts and humanities data workflows on the digital availability of cultural heritage data. Domain-specific research infrastructures like DARIAH are expected to play a key role in aligning the development of the European Data Space for Cultural Heritage with the Social Sciences and Humanities cluster of the EOSC.

1.2.2. Belgium

For the research data landscape in Belgium we look at three research funding organisations: BELSPO (federal research institutions), FWO (Dutch-speaking research institutions), and F.R.S-FNRS (French-speaking research institutions), and how DARIAH is organised in Belgium.

The federal Belgian Science Policy Office (BELSPO) introduced, through its funding schemes (e.g. BRAIN 2.0, 2018–2023), the gradual uptake of the obligation that all funded projects need to follow open data principles. Projects must use BELSPO's official data management plan (DMP, either through DMPonline or using the available Word-template) when applying; this is also taken into account by international evaluators (an 'Exceptional' rating is given when 'there is an excellent data management plan [established] in line with the highest standards, to enable easy re-use of all data'.)

- www.belspo.be/belspo/openscience/opendata_DMPintro_en.stm
- www.belspo.be/belspo/openscience/doc/BELSPO_DMPtemplate.docx

[3] Source: ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

In Flanders, the Dutch-speaking part of Belgium, the Research Foundation – Flanders (FWO) stimulates and financially supports fundamental scientific research, strategic basic research, clinical scientific research, the purchase of large-scale and medium--scale research infrastructure, and the management of large computing projects.

- www.fwo.be/en/the-fwo/research-policy/data-management-plan/

The Fund for Scientific Research (F.R.S-FNRS) has very similar goals and tasks to the FWO, but for the French-speaking community. Moreover, it animates the Data Ambassadors' Network (Réseau de Data Ambassadors de la Fédération Wallonie-Bruxelles).

BELSPO, FWO, and F.R.S-FNRS, as well as most universities and scientific institutions, all make use of DMP online in a shared, pan-Belgian instance. Researchers can log in with ORCID and view, use, and create their own DMPs.

Having a look specifically at the field of Arts and Humanities, our natural starting point is DARIAH-BE, a founding member of DARIAH-EU. DARIAH-BE consists of DARIAH--FED (federal scientific institutions), DARIAH-FWB (Fédération Wallonie-Bruxelles), and DARIAH-VL (Flanders). There are no specific DMP guidelines set by DARIAH-BE, but according to their affiliation, research projects need to follow the guidelines of their funding bodies. The interdisciplinary approach, which is so important for research, is embodied in Flanders by the creation of Clariah-Flanders, and at the federal level, by the participation of some research institutions in E-RIHS, through the KIK-IRPA created HESCIDA initiative. This also means that research data management needs to take into account the often very different sorts of research data created within the same project. An example would be a combination of metadata about a painting and its creator – high-resolution images, including infrared reflectography and macroXRF; analytical data about pigments; and more. The State Archives of Belgium participate through SODHA, which is a member of CESSDA, the Consortium of European Social Science Data Archives. In the same way that KIK-IRPA is a go-between for DARIAH and E-RIHS (with very heterogeneous research data), the Belgian State Archives play a connecting role between CESSDA and DARIAH (Social Sciences meet Arts and Humanities).

1.2.3. France

The Recherche Data Gouv project, led by the French Ministry of Higher Education and Research, aims to be France's central, federated platform for research data. It focuses on data curation, data publishing, sharing, citing, and reuse. The Dataverse-enhanced data repository Data INRAE, managed by the French INRAE - Institut national de recherche pour l'agriculture, l'alimentation et l'environnement is being expanded[4] in order

[4] Time of writing: June 2022.

to become France's cross-disciplinary research data facility, so that all French higher education institutions and researchers across all disciplines will shift from their own institutional data repositories to Recherche Data Gouv. One more central, interdisciplinary actor for research data management is OPIDoR - Optimiser le Partage et l'Interopérabilité des Données de la Recherche, which offers services in the areas of persistent identifiers (PIDs and DOIs) and data management plans. Universities, research centres, and funding institutions can make data management plan templates available on the dmponline-enhanced facility DMP OPIDoR, so that researchers can create their own data management plans.

At a policy level, in 2021, the Agence nationale de la recherche (ANR - National Research Agency) adopted the *Practical Guide to the International Alignment of Research Data Management - Extended Edition* – originally issued by the Brussels-based Science Europe think-tank – making it compulsory in France to utilise several research data management procedures (concerning DMP, FAIR, and trustworthy repositories). Training resources have been made freely available on the website DoRANum (Données de la Recherche: Apprentissage Numérique).

Specifically devoted to (digital) Humanities is Huma-Num, a French DARIAH and CLARIN chapter which offers online services, tools, policies, and documentation. Moreover, Huma-Num hosts several research project websites where research data can be found. The website Adopia - Atlas digital onomastique de la péninsule ibérique antique makes, for instance, archaeological and GIS data available.

1.2.4. Germany

Against the backdrop of the rise of data driven and digital research, there are an increasing number of calls for national approaches to bundle together the initially highly discipline-specific individual approaches to RDM into an important (interdisciplinary) infrastructural topic.[5] Big research organisations (e.g. the Schwerpunktinitiative Digitale Information, started 2008, published Guidelines 2010[6]) and funders like the German Science Foundation (DFG) took up the topic (*Gute Wissenschaftliche Praxis*, 1998; *Leitlinie Forschungsdaten*, 2015; *GWK Kommission 'Zukunft der Informationsinfrastruktur (KII)'*, 2011) and underlined the importance of research data as being national cultural heritage. Taken up

[5] More historical context given in, e.g., RfII 2016, Wuttke et al. 2021 (Wuttke, Ulrike, Heike Neuroth, Laura Rothfritz, Janine Straka, Miriam Zeunert, Carsten Schneemann, Niklas Hartmann, and Ina Radtke. 'Umfeldanalyse zum Aufbau einer neuen Datenkultur in Brandenburg: Forschungsdatenmanagement in Brandenburg (FDM-BB)'. Application/pdf. Potsdam: Universitätsverlag Potsdam, 2021. doi.org/10.25932/PUBLISHUP-48090), also Büttner et al 2011. opus.kobv.de/fhpotsdam/volltexte/2011/241/pdf/HandbuchForschungsdatenmanagement.pdf

[6] Allianz der deutschen Wissenschaftsorganisationen (2010): Grundsätze zum Umgang mit Forschungsdaten, 2 p. doi.org/10.2312/ALLIANZOA.019

as a policy area of national importance (Wissenschaftsrat 2012), the Council for Scientific Information Infrastructures (Rat für Informationsinfrastrukturen – RfII) was established in 2013. Based on the analyses and recommendations of the RfII (most importantly, RfII 2016, Leistung aus Vielfalt[7]), the outline for a national RDM infrastructure initiative (Nationale Forschungsdateninfrastruktur – NFDI) was developed further in a community-driven, participatory approach. The NFDI is envisaged as being a coordinated, distributed information infrastructure which forms the backbone of FAIR German RDM, and as a link to the EOSC. During its initial phase, it is being financed by both the national and federal German governments. Several discipline-specific consortia have already been established or are about to be established which will also cooperate with each other.[8] Some of these will be introduced in the following sections.

NFDI4Culture[9] (the Consortium for Research Data on Material and Immaterial Cultural Heritage) was approved in 2019. It aims to 'establish a needs-based infrastructure for research data which serves our community of interest, ranging from architecture, art history and musicology to theatre, dance, film and media studies'[10]. NFDI4Culture focuses on digital representations of culture and cultural objects (such as 2D digital copies of paintings, photographs, and drawings; digital 3D models of culturally and historically significant buildings; monuments; audiovisual data from music, film, and stage performances, etc.), as well as metadata, annotations, and other data obtained through research on the cultural object. The consortium not only addresses a wide range of research data and disciplines, it also consists of a highly diverse group of partners.

The goals of NFD4ICulture are,

- sustainable and long-term accessibility to cultural heritage data;
- to improve the discoverability, interoperability, and reusability of cultural heritage data;
- the provision of training in data literacy and code literacy and their professionalisation;
- promoting knowledge sharing, and enabling innovation.

NFDI4Culture is organised into different task areas which work towards the goal of finding solutions and establishing services in close cooperation with the communities. Among the first outputs from several community initiatives, events, workshops, etc., are the Culture Knowledge Graph, the Radar4Culture repository, and the Culture Helpdesk[11].

[7] Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen 2016, 160 S. English Version: Performance through Diversity rfii.de/download/rfii-recommendations-2016-performance-through-diversity/

[8] www.nfdi.de/consortia/?lang=en

[9] Webseite: Kammerer, Dietmar, Schrade Torsten, und Stellmacher Martha. 2021. 'NFDI4Culture: Konsortium für Forschungsdaten Zu Materiellen Und Immateriellen Kulturgütern'. *Bausteine Forschungsdatenmanagement*, Nr. 2 (Juli). Germany: 23-33. doi.org/10.17192/bfdm.2021.2.8341.

[10] nfdi4culture.de/index.html

[11] nfdi4culture.de/what-we-do/services.html

Text+[12] (a consortium for language- and text-based research data infrastructures): Text+ was established in the second round of the NFDI and took on its work in October 2021. It aims to 'develop a research data infrastructure for humanities disciplines and beyond whose primary research focus is on language and text'[13]. Text+ focuses on digital collections (such as the corpora of written or spoken material, e.g., literary texts, newspapers, and recordings), lexical resources (mainly dictionaries and encyclopaedias) and (scholarly) editions. A further task area comprises infrastructure and operations, in which FAIR principles and software services are addressed. The consortium is made up of 34 institutions and consists of a large community.

The goals of Text+ are,[14]

- the development of a distributed infrastructure for speech language and text data;
- support for researchers in the creation, re-use, and preservation of language and text data;
- to enable innovative research through easy access to research data and tools;
- to have close cooperation with the communities involved – portfolio expansion, training, and workshops.

The outputs of the German DARIAH-DE[15] and CLARIN-D[16] initiatives are incorporated in the work of the Text+ Consortium.

NFDI4Memory[17] (a consortium for history and disciplines which make use of historical data): NFDI4Memory received funding approval from the Joint Science Conference (GWK) in the DFG's third NFDI funding round in November 2022 and took on its work in March 2023. NFDI4Memory aims to 'establish systematic, sustainable links among three main categories of the producers and users of historical data: historical researchers, memory institutions (archives, libraries, museums, and collections) and information infrastructures'[18]. The task areas focus on data quality, data connectivity, data services, data literacy, and data culture[19]. The consortium consists of over 80 institutions and the large community behind it.

The goals of NFDI4Memory are,[20]

- to link research, memory institutions, and infrastructures;
- to integrate historical source criticism with data services;
- to create a network of historically oriented research communities;
- to create a knowledge order for the digital future of the past;

[12] www.text-plus.org/ (last accessed 22/06/23)
[13] Erhard Hinrichs, Peter Leinen, Alexander Geyken, Andreas Speer, & Regine Stein. 2022. *Text+: Language- and text-based Research Data Infrastructure*. Zenodo. doi.org/10.5281/zenodo.6452002
[14] www.nfdi.de/textplus/?lang=en (last accessed 22/06/23)
[15] de.dariah.eu/en/home (last accessed 22/06/23)
[16] www.clarin-d.net/en/ (last accessed 22/06/23)
[17] 4memory.de/ (last accessed 22/06/23)
[18] 4memory.de/ (last accessed 22/06/23)
[19] 4memory.de/our-task-areas/ (last accessed 22/06/23)
[20] 4memory.de/linkage/ (last accessed 22/06/23)

- to advance the analog/digital interface of historical source material and data;
- to generate standards for historical research data and sustainability;
- to promote education and citizen participation.

NFDI4Culture, Text+, and NFDI4Memory have, together with the fourth humanities consortium (NFDI4Objects[21]), published a *Memorandum of Understanding*[22] in 2019 in order to emphasise their joint efforts in collaboratively building up a national research data infrastructure (NFDI) in the Humanities and Cultural Sciences.

Below the national level, a lot of research-data management responsibility lies at the level of Germany's member states. One relevant example concerning digital history is the Historisches Datenzentrum Sachsen-Anhalt, i.e., the centre for historical data service in the federal state of Saxony-Anhalt.

Many universities and non-university research institutions make concrete recommendations by instituting individual (institutional) guidelines. These are quite diverse in their lengths and how binding their requirements are. As an example, in 2018, the Leibniz-Gemeinschaft (an umbrella organisation of 97 German independent research centres, some of which are active in the Arts and Humanities) issued its *Leitlinie zum Umgang mit Forschungsdaten* (Guidelines for Managing Research Data) in order to promote cooperation among members at the level of research data.

1.2.5. Poland

Although not very formalised or formally established in Poland until relatively recently, open science has been of interest to some activists and enthusiasts for a long time. The knowledge and know-how surrounding open access and research data management on the national stage has been, for a long time, developed by open science enthusiasts, including librarians, researchers, and some civil service officials. The emergence of this community can, for instance, be acknowledged through the fact that Open Access Week has been celebrated in Poland since 2010 and has, for years, been coordinated by Bożena Bednarek-Michalska and the EBIB (a professional library service). Bednarek-Michalska was one of the key figures in the early adoption of open science and has been an involved activist who ought to be mentioned. As a curator and librarian at the Nicolaus Copernicus University Library in Toruń (where she twice served as vicedirector) she also ran the portal Uwonij Naukę[23], which is about open access

[21] www.nfdi4objects.net/index.php/en/nfdi4objects-english (last accessed 22/06/23)

[22] Sabine Brünger-Weilandt, Kai-Christian Bruhn, Alexandra Busch, Erhard W. Hinrichs, Wolfram Horstmann, Martin Grötschel, Johannes Paulmann, Philipp von Rummel, Eva Schlotheuber, Dörte Schmidt, Torsten Schrade, & Simon Holger. (2019). *Memorandum of Understanding by NFDI initiatives from the humanities and cultural studies*. Zenodo. doi.org/10.5281/zenodo.3265763

[23] uwolnijnauke.pl/

initiatives. This allowed the community to be more prepared for the top-down research data measures which have come about in recent years.

In 2008, as part of the European University Association, 43 higher education institutions from Poland and the Conference of Rectors of Academic Schools in Poland participated in issuing recommendations prepared by the EUA Working Group on Open Access, in which the need for repositories was stressed.[24]

The early open science recommendations and directives in Poland were mainly communicated through open letters by the ministry and scientific communities.[25] Formally, one of the greatest developments to have increased the number of discussions, meetings, training events, and general awareness concerning the management of research data came with the requirement (formed by the National Science Centre and announced in April 2019) that all submitted proposals within certain calls should include a data management plan.[26] The proposal included only a short version of the DMP, but it still encouraged researchers and administrative staff to reflect on topics related to opening up research results. Importantly, all projects financed by the NCN (Narodowe Centrum Nauki) in calls announced on 16 June 2020 or later have been required to publish their findings in open access.[27]

While the importance of RDM has been growing (also within the Arts and Humanities), the Interdisciplinary Center for Mathematical and Computational Modeling (ICM) at the University of Warsaw launched three new Polish data repositories in 2020 as result of the Disciplinary Open Research Data Repositories (Dziedzinowe Repozytoria Danych Badawczych) project. ICM caters to various disciplines: RepOD (Repozytorium Otwartych Danych – Repository for Open Data), RDS (Repozytorium Danych Społecznych – Social Data Repository), and MX-RDR (Macromolecular Xtallography Raw Data Repository). It is also ICM who runs the Library of Science (Biblioteka Nauki – an open access platform). The Institute of Literary Research of the Polish Academy of Sciences, and, in particular, the Digital Humanities Centre and the Polish Literary Bibliography, are involved in data acquisition for the TRIPLE project, in which the GOTRIPLE platform was created.

[24] www.openaire.eu/os-poland

[25] See, e.g., Stanowisko Prezydium PAN z dnia 3 listopada 2009 roku w sprawie Open Access; Wspólne Stanowisko Prezydium KRASP i Prezydium PAN z dnia 5 lipca 2013 r. w sprawie zasad otwartego dostępu do treści publikacji naukowych i edukacyjnych; Wytyczne Ministra Nauki i Szkolnictwa Wyższego z 23 października 2015 'Kierunki rozwoju otwartego dostępu do publikacji i wyników badań naukowych w Polsce'; List Ministra Nauki i Szkolnictwa Wyższego z dnia 10 lutego 2017 r. w sprawie otwartego dostępu do publikacji naukowych i otwartej nauki; Raport MNiSW nt. realizacji polityki otwartego dostępu do publikacji naukowych w latach 2015–2017 z marca 2018.

[26] www.ncn.gov.pl/sites/default/files/pliki/2019_04_03_pismo_dyrektora_NCN_zarzadzanie_danymi_naukowymi.pdf

[27] www.openaire.eu/os-poland

Importantly, in 2021, the DARIAH.Lab project began, bringing together colleagues from 15 Polish institutions already involved in DARIAH-PL (the Polish DARIAH consortium) and the Poznan University of Technology. It is the biggest infrastructural project for the Arts and Humanities ever implemented in Poland. The laboratories making up the project include the Source Laboratory, Automatic Enrichment Laboratory, Supervised Semantic Discovery Laboratory, Intelligent Analysis and Interpretation Laboratory, and Advanced Visualisation Laboratory. The built infrastructure will be available and shared with Polish and international projects.

Following up on the *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information*, the Polish ministry introduced a relevant, data-related act on 11 August 2021. However, the implementation of the act is still in its early stages and it remains to be seen what the exact consequences for RDM will be.

More highlights of open science initiatives in Poland (not only data-specific) can be found at www.openaire.eu/os-poland.

1.2.6. Switzerland

The Swiss National Science Foundation (SNSF) supports the implementation of open research principles. Since October 2017, it has been mandatory to submit a data management plan for most funding instruments. Similarly, the SNSF expects that data prod- uced during research work will, subsequently, be publicly accessible, provided that no legal, ethical, copyright, or other issues prevent this. The same applies for the Swiss Academy of Humanities and Social Sciences (SAGW), which published a paper summarising its open science policy.[28]

In January 2020, the State Secretariat for Education, Research, and Innovation (SERI) commissioned swissuniversities to develop the *Swiss National Open Research Data Strategy*, to prepare an associated action plan, and to write a background report. The developed ORD (open research data) strategy formulates various principles, among others, stating that the FAIR principles must be applied and that research data should be as open as possible and interoperable. But it also contains additional objectives; for example, in order to avoid the development of overlapping infrastructures and services within the country, there should be coordinated governance, while skills development and best practices should also be shared among researchers. In the action plan, which covers the years 2022 to 2028, four focus areas were defined which were then further

[28] Beat Immenhauser. 2019. *Open Science Policy of the Swiss Academy of Humanities and Social Sciences*. Zenodo. doi.org/10.5281/zenodo.2634242.

structured into action lines. There have been – and still are – open calls by swissuniversities to support and develop open access to publications, ORD practices and community building, and to promote data stewardship and ORD specialists at all higher education and research institutions.[29] The latter resulted in 25 higher education institutions in Switzerland receiving funding to build up data stewardship in the respective institution.[30]

Concerning the development of national infrastructures in the field of Arts and Humanities which have a mission to archive research data for the long-term and to support their community in all matters relating to RDM, two infrastructures (which were pre-selected based on the 2019 Swiss Roadmap for Research Infrastructures) were invited to participate in a closed funding call by the SNSF: they were the Swiss Center of Expertise in the Social Sciences (FORS) and the Swiss National Data and Service Center for the Humanities (DaSCH). Both institutions succeeded in securing funding for the period 2021–2024. The selection of new infrastructure projects using multiple stages of evaluation is currently ongoing. In 2023 a new roadmap will be published by SERI and a new call will be released by the SNSF for the funding period 2025–2028. Swiss national infrastructures are thus subject to the same conditions as usual research projects, with funding for four-year periods. Besides the two already mentioned discipline-specific national infrastructures which are relevant to the Social Sciences and Humanities, several general purpose and usually uncurated repositories exist in Switzerland, with Zenodo being the most well-known, which is operated by the European Organisation for Nuclear Research (CERN).

European infrastructures are also important for Swiss researchers. As early as 2018, the first national DARIAH consortium was established; and in 2021, Switzerland was able to join DARIAH as an observing member. For Switzerland to achieve full membership, the Swiss parliament had to adapt their laws. This was successfully achieved, and starting from June 2023 Switzerland will be a full DARIAH member. For Switzerland, a similar procedure applies to membership of CLARIN-EU. The consortium for CLARIN-CH was founded in 2021.

Switzerland is a country with a bottom-up organisational system, and hence it is not surprising that each higher education institution has its own structures and ways of supporting their researchers. Nevertheless, despite all the differences in the details, fundamental commonalities can be identified in the various solutions. Thus, the following short overview aims to highlight this common ground. The provided links to the

[29] As of April 2023, the content of the ORD calls webpage will change; more calls are expected.

[30] www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/ORD/Calls/B5.2/Genehmigte_Projekte_B52_Data_Stewardship.pdf.

institutions at the end of the subchapter will allow interested readers to explore the specific solutions.

Since April 2023, researchers who request funding from the Swiss National Science Foundation (SNSF) no longer have to submit a data management plan together with their project application, but must deliver one within six months if the project will indeed be funded. SNSF as well as several other Swiss funding bodies require open science, as far as possible, to allow public access to publications and research data to be free. Thus, researchers are forced to think about what type and what amount of data they are going to create, how they want to deal with this data during the lifetime of the project, and in which repository they want to deposit the data afterwards. The SNSF provides researchers with a list of accepted repositories, subdividing them into generalist, discipline-specific, and institutional categories. The funding bodies' request triggered the development of structures and services at Swiss higher education institutions, which support researchers on questions concerning, for example, how to write a DMP, which tools to use during the lifetime of the project, and when and how to deposit the data. A rather common approach is that university libraries, IT services, and a third variously termed facility which consists of staff with discipline-specific knowledge are involved. The respective facilities in university libraries usually provide basic general information, offer help concerning DMPs, and redirect researchers to other involved divisions. IT services give advice on technical questions such as storage solutions during the lifetime of the project or about special software for RDM, and provide servers if necessary, and sometimes also hosting. The third division comprises data stewards and research consultants who – depending on the institution – take more or less active roles in research projects and in preparing data for data repositories. In some of the institutions, for example, the universities of Lausanne and Basel, there are only two divisions involved because research and infrastructure support is combined. In contrast, the University of Zurich has numerous agencies involved throughout the whole process. For a description of how this structure has developed in the case of digital editions, see Malits 2020. Another important service is to provide students at all levels with the necessary digital skills – such courses exist at several Swiss universities.

- Università della Svizzera italiana
  RDM: www.usi.ch/it/universita/info/srit/research-data-management-service
  Research support: www.usi.ch/it/universita/info/srit
- University of Basel
  RDM: researchdata.unibas.ch/en/
  Research & Infrastructure Support: rise.unibas.ch/en/

- University of Fribourg
  RDM: www.unifr.ch/researcher/en/openscience/
  Research Support: www.unifr.ch/researcher/en/contact/promrech.html
- University of Geneva
  RDM: www.unige.ch/recherche/en/policies/
  Research & Infrastructure Support: www.unige.ch/eresearch/en/services/hedera/
  Repositories: olos.swiss/, www.unige.ch/eresearch/en/services/yareta/
- University of Lausanne
  Information resources and archives: www.unil.ch/uniris/home.html
  Research & Infrastructure Support: www.unil.ch/ci/dcsr-en
- University of Zurich
  Open Science Services: University Library Zurich - UZH
  Research and Publish: www.ub.uzh.ch/en/wissenschaftlich-arbeiten.html
  Service and Support for Science IT (S3IT):
  www.zi.uzh.ch/en/teaching-and-research/science-it/
  Linguistic Research Infrastructure (LiRI): www.liri.uzh.ch/en.html
  Center for Digital Editions: www.zde.uzh.ch/en/projects.html
- Swiss National repositories
  SWISSUBase of the Swiss Centre of Expertise in the Social Sciences (FORS):
  www.swissubase.ch/en/
  The DSP of the Swiss National Data and Service Center for the Humanities (DaSCH): www.dasch.swiss/

The perspectives related above, showcase the diverse translation and implementation routes of the open and FAIR research mandates in five European countries. The GiLab book format allows this chapter to be dynamically expanded, either to include new countries or by complementing and eventually correcting the information above as these policies evolve. To keep it as rich and up-to-date as possible, we hereby invite our readers to enter into co-authorship with us.

2. THE INSTITUTIONALISATION OF NEW DATA SUPPORT ROLES ACROSS EUROPE – A CALL FOR AN EXCHANGE OF EXPERIENCES AND FOR SOLIDIFYING GOOD PRACTICES ALONG ARTS AND HUMANITIES LINES, ACROSS EUROPE

Femmy Admiraal, Marta Błaszczyńska, Mirjam Blümm, Vera Chiquet, Rita Gautschy, Peter Gietz, Erzsébet Tóth-Czifra, Maria Vivas-Romero, Ulrike Wuttke

The voices of our Working Group

One of the motivations which brought our Working Group (WG) to life was that we saw a strong need to build a European knowledge hub for researchers, cultural heritage professionals, and new data support professionals in the Arts and Humanities, to reduce the gaps between the nascent European and national FAIR and open data mandates, and the research realities in the disciplines.

Keeping in mind the terra incognita nature of the endeavour and the fact that no canonical syllabi exist for these newly emerging research support roles, in this chapter we move away from the conventions of scholarly writing and instead provide a collection of personal reflections and testimonies from our Working Group members – allowing for a sneak peek 'behind the scenes'.

We, the WG members, share our experiences and insights into the special flavours of support for arts and humanities research-data management in different national and institutional contexts; how certain institutions – our authors' workplaces – developed capacities for data support roles and how we found our way into them; what daily challenges face all of us, regardless of the institutional specificities or national legislation; and what the early take-aways are and how such first-generation experiences can shape the future of these profession(s).

Maria Vivas-Romero, Faculty Data Steward and Secretary of the Platform for Research Ethics and Integrity, Maastricht University

As a former ethnographer during my training as a scholar and coming from the global South, I became really interested in knowing what caused the disparities in access to knowledge and education. Luckily, in 2019, after realising it would be hard to obtain funding to work solely on this area, I was hired by the Maastricht University Community for

Data-Driven Insights. Indeed, in the summer of 2019, the Community for Data-Driven Insights[31] placed additional capacity on data stewardship to help accomplish the open science and FAIR agenda in all scientific fields. In this context, I was outsourced to the Faculty of Arts and Social Sciences (FASoS), where I had previously conducted part of my PhD studies on care and migration. Open science and FAIR were high on the agenda of our former rector. In 2019 she signed the DORA declaration and committed to making all research output a democratic resource for all researchers and citizens everywhere. As a former researcher and ethnographer, I became a FAIR translator and communicator for FASoS. Here I reflect on this process.

The Community for Data-Driven Insights gave the data stewards the mission of drafting and implementing a FAIR action plan. I discuss below how, as a data steward, I engaged with a policy working group which drafted the *FAIR Action Plan* for the Faculty of Arts and Social Sciences. The group of experts was very diverse and included the dean of research, the cluster coordinator, the information manager, and the communication head. However, as a former ethnographer, I became the only actor involved who clearly understood that the data produced in the faculty, which emerged from grounded approaches, could not always become a digital object and thus entirely FAIR. This non-digital component is a challenge for Social Sciences and Humanities. While reading Wilkinsons et al. (2016), the call seems to be to make research output digital and available both for humans and machines, both now and in the never-ending future. In this context, the aim of the *FAIR Faculty Action Plan* became, first, to translate these principles into research data which included both digital and non-digital forms. Second, the aim was also to convince all researchers that, for them, there was something to be gained from this cultural change. Indeed, not all researchers are convinced that open science was a gain for the greater good of science, and some feared losing the freedom of doing science. This process has been taking place and it's still happening now.

It is divided into four phases:

- Communication and awareness surrounding FAIR principles and open science.
- Translating findability and accessibility into practice through DataVerseNL and research registration workflow; finally.
- Work to establish an automated workflow for FAIR/GDPR compliance in the Humanities and Social Sciences.

In the following two paragraphs, I will describe these phases in detail.

[31] See more at our portal at: Research Data Management and FAIR - Maastricht University.

The awareness campaign was focused on having researchers understand that it was also an advantage for them to join the open science and FAIR movement. FASoS is divided into research departments, and every department has a different structure and collects different types of data. We concluded that six types of data were used or produced by FASoS researchers: 1) interviews with experts and lay people, 2) surveys, 3) field observations and research diaries, 4) collecting data from private archives, 5) secondary analysis of (several) existing data sets, and 6) using data from existing archives and collections. Categories 1 to 5 create data which is open to FAIR considerations in principle, and which may be 'personal' data according to GDPR stipulations. After conducting this awareness campaign, the greatest commitment came from PhD researchers; and later, when founders and publishers began to request FAIR data practices, the entire community joined in.

In 2020, in the midst of the global pandemic, the faculty decided to commit to making their research data findable and accessible. To do so, we needed to look into the quality of the metadata, and the neatness and curation of our DataverseNL folder. Using DataverseNL ensured the quality of the meta-data used. I built a workflow from scratch which included drafting a DMP, personal data registration, storage of data within our local networks, and, eventually, the publication of the meta-data and/or data on DataverseNL. The latter has resulted in a significant augmentation of data management plans, and meta-data which has been published to the DataverseNL repository.

Rita Gautschy, Director, DaSCH, Switzerland

I developed the capacities for my current role as director and head of the Research Data Management Specialists Team at the Swiss National Data and Service Center for the Humanities (DaSCH) over a couple of years, and this was mainly a learn-by-doing process. I began 'on the other side' in 2015, when I became responsible for a digital, open access version of an important archaeological lexicon which comprises 20 volumes and was compiled over almost 50 years – the Lexicon Iconographicum Mythologiae Classicae (LIMC). At Basel University we had an archive of images of archaeological objects from all over the world in our basement and a technically obsolete database which was initially set up to compile an index for the printed books. We wanted to make the collected material freely accessible so it would be easily available to colleagues in countries where the printed books were not available due to their high cost. We also wanted our data to be interoperable with many other archaeological data(bases). In collaboration with colleagues from the Digital Humanities Laboratory (DHLab) at the University of Basel, a data model was created, and the existing data cleaned and imported into a system for the long-term preservation of data which was developed by

DHLab at the time. During this time I learned the basics of the Semantic Web, standards, norm data, RDF, conceptual modelling, and the different types of databases.

DaSCH was founded in 2017; in 2021 it became separate from DHLab. DaSCH is a FAIR repository for humanities data which keeps databases alive. This means that the data from research projects remain queryable directly via a generic web application. Thus, each research project has to define its own data model, or at least to modify an existing one. In 2019 I joined the DaSCH team and switched roles – from this time it was me helping and teaching researchers how to model and clean their research data, and, finally, to import these data into our system for long-term preservation. I like this role very much, as it requires a range of skills. First, at least basic programming skills are necessary in order to clean data, prepare them for import, and finally import them. Second, it allows me to get a deeper insight into different fields within the Humanities and to work in close contact with the researchers – this is important as, otherwise, the result won't be of good quality. Third, in the meantime I teach students and I am always happy if they are able to, increasingly, take over parts of my own work, for example, the data cleaning. Fourth, good social skills are also required, especially if younger researchers become very stressed because they are not on track with their project for the next intended career step. Difficult situations are likely to occur if there is a mismatch between the aims of the project (e.g. sophisticated application functionalities) and the available skills within the project team. In my experience the average digital skill level of humanities researchers in Switzerland is still rather low. This everyday experience has made me aware of the need to offer appropriate courses in our curricula at a much lower level than is currently the case (bachelor vs. PhD programmes).

The main lessons I have learned so far are that no one-solution-serves-all, and that the perfect data model, from a modelling point of view, may not be the best solution and may be much too complicated for everyday use – the fine art is to find the golden mean. My path is probably a typical example of how someone slipped into this data support role in the first half of the 2010s. In the meantime the landscape has changed.

Marta Błaszczyńska, Vice-Director (Open Science) of the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, Poland

Although not a data steward by title, I thought I'd throw in a couple of reflections about the increasing amount of data-related activities which have been included in my open science role at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN) since I started working here in May 2019. While my initial tasks, when first employed as an open science officer within the institution, focused more on generally

working toward the implementation and advocacy of open access (collaboration with the institutional publishing house, supporting the editorial teams for IBL PAN's journals etc.), the challenges have been becoming increasingly data-driven. This, of course, is unsurprising considering that higher level open science is impossible without reflections about opening up research data.

In 2019 I was approached about creating an institutional template for a research data management plan. For me, this was one of the first steps in truly comprehending the importance of proper data handling in the Humanities. Slowly understanding the basics, I increasingly perceived RDM as a scholarly activity, and a well-developed DMP as the result of deep scholarly reflection. In order to develop my skills, I participated in training provided by the Polish National Science Centre in Kraków in October 2019 and the DESIR (DARIAH ERIC Sustainability Refined) winter school: Shaping New Approaches to Data Management in Arts and Humanities. It was also this event which later allowed me to get involved in the Research Data Management Working Group!

Data was also a big part of the *Open Science Policy* created for the IBL PAN by the Digital Humanities Centre. While drafting the document we had several consultations with the institutional publishing house, research support centre, library, journal editorial boards, and archives. Together with my colleagues Dr Maciej Maryl and Mateusz Franczak we wished to create an open, collaborative process which would reflect the openness of the policy we were formulating. One of the appendices to the *Open Science Policy* contains practical advice on preparing a data management research plan, stressing that the guidelines cannot be taken universally and need to be adapted to each project. Moreover, our *Open Science Policy* has a whole section devoted to research data, explaining the importance of data storage for data reuse in the future, clarifying the FAIR principles, explaining the importance of the institutional repository, etc.

An important moment for research-data literary studies came at the same time as the series of three workshops which we organised for Polish scholars, in which we developed a taxonomy of data in literary studies. I explore this theme in more detail in Chapter 3.1.4.

The growth of the open science team which I now have the pleasure of managing within IBL PAN allowed us to diversify our efforts and for me to concentrate my interests on research data, but – as it often does – the real issue lies in funding. Since several of my teammates are employed within projects and are not permanent staff members, delegating tasks to them which are related to ongoing institutional support is often impossible. My wish for the future is to gain sustainable sources of funding so we could be even more proactive in our open science endeavours.

Ulrike Wuttke, Professor of Library Science, University of Applied Sciences, Potsdam, Germany

I was the Interim Professor for Library and Information Technology and Digital Services in the Department of Information Sciences at the University of Applied Sciences, Potsdam (FHP), Germany, from 2020, before becoming Professor of Library Science in 2023. My teaching and research interests include digital research infrastructures, research data management, digital humanities, scholarly communication, and open science. I have contributed to various national and international networks in these areas, such as the Working Group Data Centres in the Digital Humanities im deutschsprachigen Raum (DHd), and the Culture Steering Board of NFDI4Culture.

In my current position I mainly teach (among other subjects) research data management at bachelor and master's levels at the University of Applied Sciences, Potsdam. I also contribute to the RDM-Working Group of my institution, which has only recently released the first institutional guidelines on RDM (*Forschungsdatenleitlinie der Fachhochschule Potsdam* 2021) and signed DORA as well as the Berlin Declaration on Open Access. Before this, I was task leader within the FDM-BB project (FHP), the RDMO project (FHP), PARTHENOS project (FHP), scientific coordinator for the AGATE-project (Union of the German Academies of Sciences and Humanities), and worked for the Humanities Data Center (Göttingen Academy of Sciences and Humanities) and the Göttingen eResearch Alliance (Göttingen State and University Library).

What unites these projects and my position is a focus on Digital Humanities and Research Data Management as well as an emphasis on training and education. Coming from a humanities background (PhD in Medieval Studies) with an additional Master's in Library and Information Science (LIS), awareness raising and capacity building within the communities are, for me, very satisfying but sometimes also highly challenging aspects of RDM. I have given various training sessions about writing data management plans and on digital research methods, and still do so, while also using this experience to contribute to the *Train-the-Trainer Concept on Research Data Management*, which is a very helpful resource for learning about RDM as well as conducting RDM training. I have also published about aspects of RDM on my website blog, as well as slides and other training materials on Zenodo (you can find a list on my website).

Femmy Admiraal, Data Station Manager, Humanities, KNAW-DANS, the Netherlands

Trained as an anthropologist and linguist, I started working on language documentation for my PhD. This project resulted in an extensive documentation of the endangered language called Baure – this dataset is archived[32] at The Language Archive, Nijmegen. Working at KNAW-DANS since 2017, I am an expert in research-data management and FAIR data. I have applied this expertise within the European infrastructures DARIAH and CLARIN, as well as in various research projects, such as PARTHENOS and Polifonia.

KNAW-DANS is the Dutch national centre of expertise and repository for data, based in The Hague. With nearly 190,000 datasets and a staff of 60, DANS is one of the major repositories in Europe. Our repository service is based on technology from the Dataverse platform, an open source platform developed by an international community led by Harvard University. We have four different data stations, each of which serves a particular domain: Social Sciences and Humanities, Archaeology, Life and Medical Sciences, and Nature and Technical Sciences. Via a user interface, any researcher can deposit a dataset to one of the data stations, which is then approved by one of our data managers before it gets published. After publication, DANS takes care of all issues involving long term preservation, such as data curation, file format migration, etc. Our repository services are certified by a Core Trust Seal and a Nestor seal.

As a data station manager for the Humanities, I support humanities scholars in various ways in their efforts to provide sustainable access to their research data. I am involved in a number of research projects, including the OH-SMArt project, where we aim to create a pipeline between a national cloud service and the DANS Data Station SSH, specifically for Oral History data and the Tales from the Drug Closet project, which aims to build an application via which (recreational) drug users can record their stories – these, then, contribute to research as well as to the public debate on safe drug use. In addition, I am the chief data officer for CLARIAH and work together with the other CLARIAH repositories to build a FAIR dataset register which serves as a large catalogue containing references to as many humanities datasets in the Netherlands as possible.

In the Netherlands, over the past decade or so, an increasing number of data stewards have been appointed to research institutions. This was stimulated in particular by the Dutch Research Council (NWO), when the Implementation Plan Investments Digital Research Infrastructure call was launched in 2019. The call consists of four pillars:

- Local digital competence centres to be hosted in research institutions;

[32] Femmy Admiraal, Franziska Riedel, Swintha Danielsen, Lena Terhart, Baptista, and Wallin. (1749–2010). Collection 'Baure'. The Language Archive. hdl.handle.net/1839/c1b94334-0fc0-4658-afbb-1b13f945eda7. (Accessed 2022-06-23)

- Thematic digital competence centres to be hosted at relevant institutions, but serving an entire domain;
- Investment in eScience to support collaborative research projects;
- Computer facilities to make use of either the Snellius supercomputer or to apply for computing time on one of the national services.

Due in large part to this call, most universities now have a dedicated team of professionals which support their researchers with data management. Often, these professionals are bound to the university library or to the various faculties. For example, at Leiden University, data stewards at the faculty level offer first-line support directly to the researchers. Acting as second-line support, the Centre for Digital Scholarship of the University Library maintains this network of data stewards, and is responsible for the implementation of research data management support services at a university-wide level.

Vera Chiquet, University of Basel, Switzerland

Trained as an art historian and sociologist, I started working on image reproductions for my PhD, where I researched photographic manipulation in the media (see, *Fake Fotos*, doi.org/10.1515/9783839441442-005). During my research, I collected a lot of digital research data, mainly digital photographs of analogue magazines, artistic drafts, and case studies. I wanted to store these valuable documents – my research data – in a way which others could access them. That's why I turned to the Digital Humanities Lab at the University of Basel.

It was not because of technical issues but due to an unresolvable legal issue at that time which we, unfortunately, had to decide not to pursue this further, but this led me to start work there in 2017. I got acquainted with digital long-term archiving; the Semantic Web and RDF; PIDs; standards and computational photography, like RTI (joining a start-up which grew from the former imaging and media lab); and photogrammetry.

I joined DaSCH during its start-up phase and worked closely with CHI and their collection, transferring it to a graph database. Besides digital project management, I worked for the DARIAH-CH Consortium, which is coordinated by DaSCH, until I stepped into the position of interim head of the Department Professorship Digital Humanities at the University of Basel. Here I teach about theoretical issues concerning digital knowledge organisation and its history, and offer hands-on courses about digitising analogue collections. Together with the start-up Virtual Culture, founded in 2021, I am busy with the daily business of answering questions about RDM, especially for CHIs. Here it is important, probably even more so than in university research projects, to implement

pragmatic solutions and to be as open as possible but also as fast as possible in implementation. These are small but efficient projects which, for example, want to make intangible cultural heritage available digitally – for instance, the webarchive of the Schnitzelbänke.

Mirjam Blümm, Professor for eScience and Research Data Management, Cologne University of Applied Sciences (TH Köln), Germany

In 2018, TH Köln was the first university of applied sciences in Germany to advertise a professorship for eScience and research data management. I was chosen for the position not least because of my previous work as co-head in the DARIAH-DE coordination office. Here I've had the opportunity to gather a lot of experience with different kinds of research data, questions of accessibility and interoperability, and sustainable research (data) infrastructure. Moreover, I have been able to profit from the exchange within the lively community at DARIAH-DE and DARIAH-EU.

The professorship is divided equally between the Institute of Information Science and the Institute of Computer Science. In the latter, from my experience in digital humanities and project management, I contribute to the teaching of future computer scientists in scientific research and writing, computer ethics, research methods, and some aspects of social computing at the BA and MA levels. In the former – my information science responsibility – I place a stronger focus on research data management. I mainly teach research data management and digital research infrastructure in BA and MA courses on library and information science, and data and information sciences.

In 2021, together with some colleagues, I started a certificate course about research data management[33], which is aimed at employees from active research and science-related infrastructure areas and designed as training. It comprises modules on the research cycle in various disciplines, open science, RDM consulting, technical infrastructure, data management, legal aspects, etc.

I also do research on the topic of RDM. In the SAN-DMP[34] joint project, we investigated how research data management can be systematically supported and established at universities of applied science using the central instrument of data management plans (DMPs). Recently, we began a project called FDM@Studium.nrw. This cooperative project aims to create reusable materials for RDM in teaching.

[33] www.th-koeln.de/weiterbildung/zertifikatskurs-forschungsdatenmanagement_82048.php (last accessed 22/06/24)
[34] www.th-koeln.de/en/information-science-and-communication-studies/san-dmp_90689.php (last accessed 22/06/24)

Erzsébet Tóth-Czifra, Open Science Officer, DARIAH, Germany

It may be that, in my case, it all started with not having institutional access to MAXQDA, a piece of qualitative annotation and analysis software, including (and also locking away) all the data which my Western colleagues have been creating and curating within the virtual environment of the tool. As a Hungarian researcher in linguistics, I come from an academic environment where visiting university libraries in foreign countries to study cultural artefacts or to learn about the latest research findings is still a common scholarly practice. So this was my first rather direct experience with the disparities in access to knowledge. On the lucky side, in 2013, when I started my PhD in Cognitive and Cultural Linguistics, the Hungarian National Reference Corpus was already in place and it opened up new, data and usage driven dimensions for conducting research into linguistics and working with big volumes of morphologically annotated corpus data. That said, it became clear early on how much the availability (or non-availability) of infrastructure, research tools, and research data defines the scope of what kinds of scholarship are possible to access and what kinds of research questions can be asked.

Discovering open access and open science, and their potential to mitigate such disparities in access to knowledge sparked my curiosity, and I was tempted into learning more about these paradigms and associated innovations. After my PhD, I became a member of a startup team called ScienceOpen, which was dedicated to the development of a research discovery platform for scientific publications. Here I had the chance to learn the basics of XML editing and its standards, metadata conversion, the value and diversity of PIDs, about scientometrics, and also the prestige politics behind all of these and open access's attempts to change these for the better.

In 2018, I was lucky enough to be given the chance, as DARIAH's open science officer, to explore and even foster the means of open research culture as they specifically make sense in the Arts and Humanities domain. As an organisation which connects several hundred scholars and dozens of research facilities, and has tools and services in 20 European countries, facilitating access to resources to make scholars' lives easier is in the DNA of DARIAH. The decision to follow and eventually shape rising open science policies and infrastructure in Europe, and to explicitly create a dedicated position for this area, came in early 2018. This included promoting innovation through new networks and collaborations, and also took the shape of Horizon2020 projects such as PARTHENOS, HIRMEOS, OpenAIRE Advance, and SSHOC; creating support and training materials for the uptake of new research and publication practices, often at the request of national DARIAHs; setting up and running the DARIAH Open Science Services Suite; having a strong voice for Arts and Humanities in European policy debates around

open science and FAIR; and, finally, doing meta-research (such as this one: zenodo.org/record/4922538#.YzHYVbRBxD8). An overview of the first three years of open science at DARIAH-EU can be found here: zenodo.org/record/5863209#.YzHVTrRBxD8.

Following the development of European policy instruments, our initially strong focus on supporting open access in publications in the Arts and Humanities has gradually shifted and diversified towards shared challenges concerning FAIR data implementation across the domain. This requires an even deeper oversight of the very diverse epistemic traditions living under the umbrella term of 'Arts and Humanities'. Apart from constant learning, checking the literature, and gaining precious experience in order to collaborate with specific disciplinary communities (such as literary studies in the CLS Infra project, or arts disciplines in the OS-ADM project), launching the Research Data Management Working Group was also instrumental in expanding and exchanging knowledge and in learning from those experts who form the Working Group.

Peter Gietz, Co-founder and CEO, DAASI International, Germany

When I graduated in Indology and Religious Studies, I had already been involved in DH for quite some time.This was especially in the context of the TUSTEP community, where I participated in the creation and publication of *An Epic and Puranic Bibliography*. Since I could better feed our children by doing computer programming, I increasingly got involved in computers and less in humanities research. I was part of the DFN research project on Internet database technologies (X.500 and LDAP), where I was also involved in IETF (Internet Engineering Task Force) standardisation processes; following this I worked at DANTE in Cambridge where I was responsible for the Root X.500 server.

Back in Germany, and back to the German Federal Ministry of Research (BMBF) funded DFN research project, my task as then project lead was to define a business model for a sustainable follow-up organisation, as the project had created infrastructure which was not fundable by research project schemes. Unfortunately (or fortunately?) the two organisations designated as the shareholders of a non-profit company backed out of the plan; I was told, while quite literally being slapped on the back: 'if you believe in your business model, why not create the company yourself?', which is what my wife and I actually ended up doing in 2000.

Since then I have been trying to run a company which is involved in open source identity and access management and digital humanities with the aim of providing good and meaningful jobs, and taking part in equally meaningful projects. We stand for digital sovereignty, and thus advocate for open source, open standards, and data privacy. With this mindset we have been involved in many research projects on research

infrastructures (TextGrid I-III, DARIAH-DE I_III), in dedicated DH projects (Relationen im Raum), and in projects on authentication and authorisation infrastructure (AAI) such as the D-Grid Projects, GAP-SLC, IVOM, and the EU Project AARC I-II. Since higher education institutions are at least a third of our customer base, we became experts in federated identity management, providing services to customers for a number of open source solutions, i.e., Shibboleth, SATOSA, and SimpleSAMLphp.

We also work for selected public agencies and private companies. Basically, with the flexibility of a private company, I was able to bring together my two main subjects of interest: IT infrastructures and humanities. Technically, the interoperability of the XML and the LDAP data model led to a number of interesting results, for example, projects like Relationen im Raum and TinCap. While being the CEO of my company, I had two employment terms at universities: one year at the University of Göttingen, leading the TextGrid Work package on infrastructure; and two years at the University of Heidelberg, leading the IT department of the Cluster of Excellence 'Asia and Europe'.

3. THE SPECIAL FLAVOURS OF ARTS AND HUMANITIES IN RESEARCH DATA MANAGEMENT

3.1. The lack of consensus around the notion of data

Marta Błaszczyńska, Bartłomiej Szleszyński

3.1.1. Digital scholarly editions, TEI, and FAIR-ness

In this subchapter we will further explore data-related insights into **digital scholarly editions**, **digital collections**, and digital monographs. These will act as specific case studies illustrating the relevance of in-depth data enquiries in literary studies which are, indeed, necessary to answer scholarly questions and to conduct literary research in an organised manner.

Scholarly editing (understood as the work carried out by specialists to determine and develop the best version of a literary text and create a scholarly commentary on it) is one of the important areas within the broad scope of literary research. On the other hand, the digital variant of scholarly editing, DSE (digital scholarly editions), has been evolving continuously over a long period of time, not only developing general methods of operation, but also implementing the TEI (Text Encoding Initiative) standard for marking and describing individual editions. In the context of reflecting on research data, this leads to the question: Is TEI FAIR?

3.1.2. Is TEI FAIR? (Or; How can we understand FAIR data in the context of DSE?)

The most important issues related to data in DSE are centred around ways of using the TEI standard and the question about what FAIR rules might mean in the context of data in DSE. While the approaches to DSE can vary quite a bit from project to project, it is undeniable that TEI is currently the only standard which they have in common. The authors of the report *Seeing Shapes in the Cloud: Perspectives from the Humanities on Interdisciplinary Data Integration* (Doran et al. 2022) even stated that

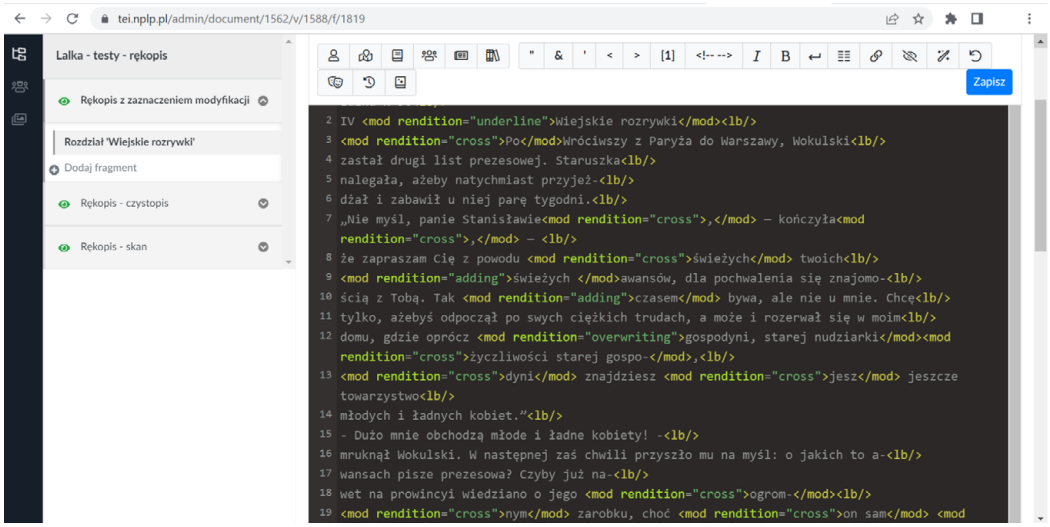
> [o]ne of the most successful implementations of a standard in the Digital Humanities community has been that of the Text Encoding Initiative (TEI). (Doran et al. 2022)

It is extremely interesting, and requires further reflection, that the reasons cited in this report for the popularity of TEI are

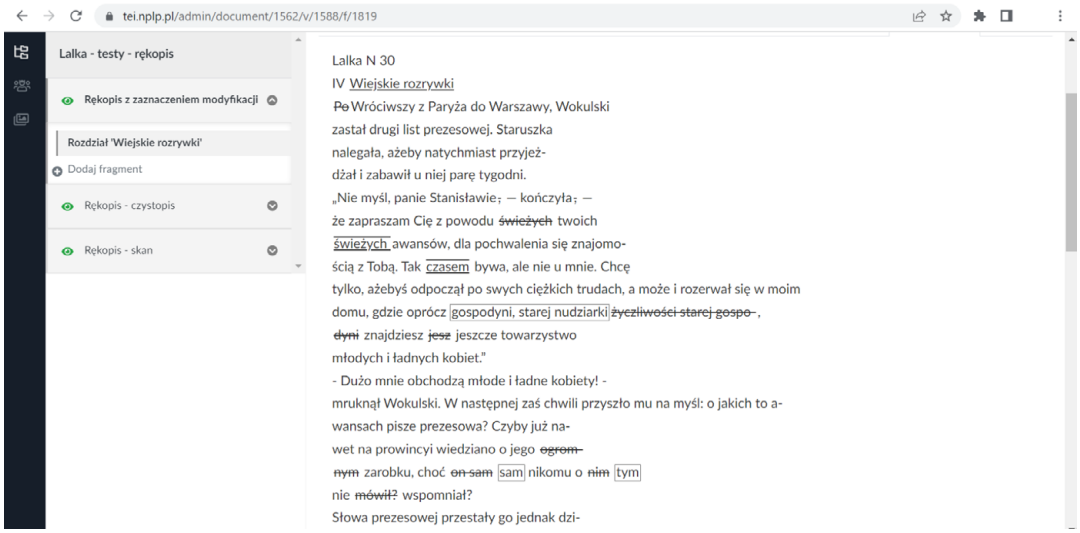
> [w]ith the exception of a few formal elements required in the header, the TEI does not require a user to mark up particular aspects of a text, only to mark up those aspects considered important in a certain way. And the number of possibilities is vast: in its most recent release, TEI P5 contained 569 different elements to choose from. (Doran et al. 2022)

The fact that 'TEI guidelines' are not 'TEI rules', might be considered both an advantage and a disadvantage. It is also worth noting that in addition to the intended flexibility of the TEI standard, there are areas where the creativity of editors must go beyond the few options described in the TEI guidelines. There are also tagsets which are not sufficient to describe some phenomena (such as TEI Drama for describing complex didaskalia in postdramatic plays) which necessitate the development of unique TEI applications for specific projects. This leads to a situation where there are multiple editions using TEI, but each in a slightly different way. This can cause incompatibility effects across different DSE projects at multiple levels. One of those levels is the ability to visualise a particular text marked with TEI in different software.

For example, when we mark up specific manuscript properties (such as deletions or additions) in the TEI editor, as illustrated below.



we want the front-end software to show these properties in the appropriate way, as illustrated here.



Figures 1 and 2. *Screenshots from the online application tei.nplp.pl*

This can only be done if every TEI markup has its own precise graphical representation.

Texts which are tagged in TEI in a different way than those provided for by the visualisation software, will not be displayed correctly – or as expected.

Even more complicated is the issue of automatically processing different texts which have been tagged in TEI in different ways (e.g. for statistical analysis of the text, its structure, or connections between elements such as people or places). The correct results will not be provided without unifying the structure of the way in which the analysed phenomena is marked. This leads to the recognition that while findability and accessibility are not a problem in DSE (as on the platform TEI.NPLP.PL, where we can easily search for texts and retrieve TEI encoded versions), interoperability and reusability may be limited to projects which use TEI markup in the same way.

It seems that if we are considering ways to integrate infrastructure for DSE (on a multinational and interdisciplinary level), it is also necessary to think about re-standardising, at least partially, the use of TEI for different projects. The first step should be to draw up a protocol of discrepancies between the different DSE projects, that is, a list of differences in the way the TEI standard is used, with recommendations for their unification.

3.1.2.2. Other types of data in DSE

DSE also contains content which can be viewed as data other than text marked in TEI: programers code for backend and frontend, descriptions of entities, and a web of connection between entities.

Sharing programming code (software) is good practice; it not only lowers the entry threshold to DSE for different institutions and researchers, but also leads to more DSEs using TEI in the same way.

Quite often, DSE contains descriptions of marked entities (people, places, organisations etc.).
In the case of the TEI.NPLP.PL platform, all descriptions are created for specific projects by professional researchers; these help in being able to better understand the edited text, but also form a set of very rich and reliable data.

Again; there is no problem in finding and accessing them. They are findable and interoperable, but reusing them is limited to the same platform. As good practice, it would be helpful to connect such descriptions using VIAF (virtual international authority file), Wikidata, etc. To make this kind of data reusable outside a particular platform, sets of such descriptions from different projects could also be placed in repositories after work on the specific project ends.

DSE can be seen as a text or as a set of data, but also as a tool. A good search engine, filters, and web of connections between marked entities may allow scholars to answer different research questions. You can, for example, trace all the ways in which a person is referred to in a text. Below we can see examples of expressions from correspondence about Kazimierz Wierzyński, a Polish poet.

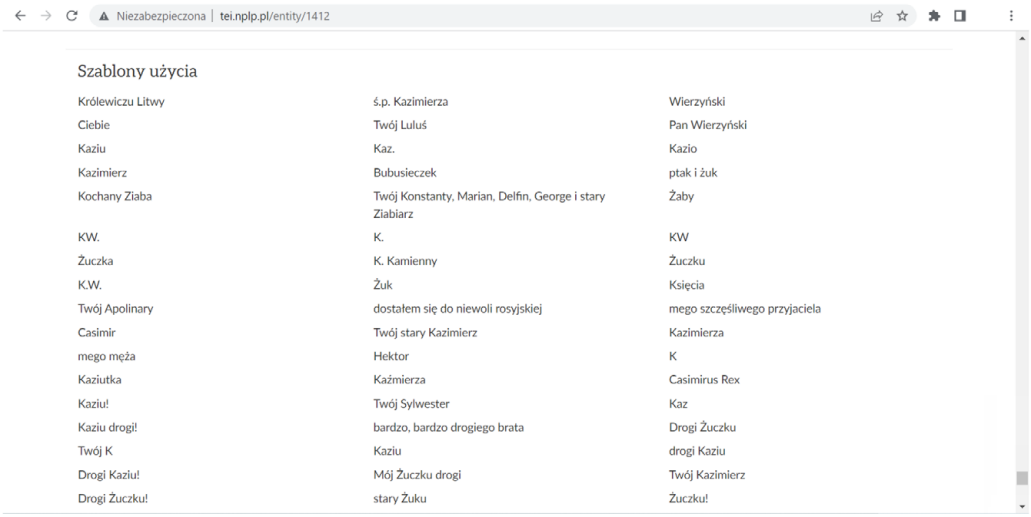


Figure 3. Screenshot from the online application tei.nplp.pl

From this perspective, data on this platform is reusable – and the more data on the platform, the more complicated research questions can be asked in relation to the data. The considerations of the authors of the *Report on the Future of Scholarly Writing in SSH* still hold true – standardisation (regarding TEI, but also for other aspects), is one of the most important challenges for DSE in the future – certainly, considering the

possibilities for FAIRification can help develop the paths it should take.[35]

It should be noted that the issues described in the above subsection are not limited to DSE in literary studies, nevertheless, it seems that they are extremely important, as editing is a highly important area of literary studies.

3.1.3. Digital monographs/collections

It is worth considering the issue of data in digital publications which we call digital collections or monographs. Examples will be given from the New Panorama of Polish Literature platform, where various scholarly digital collections have been published over many years. This is a good case study because the platform's main focus has not been on data standards but on scholarly narratives – analysing it in terms of data (and their possible FAIRification) may reveal more general challenges in this area.

Digital monograph

The 'traditional' form of a scholarly monograph in literary research is a book written on a specific subject (usually on a literary work and its authors) by one or many authors. There are several traditional elements to the scholarly monograph: the main text, bibliography, footnotes, table of contents, preface, index of names, and sometimes other indexes. Some of these, transferred into the digital environment, can be treated as data. But of course, digital monographs are much more than a simple transition of these traditional elements onto the Internet.

A good illustration of a digital monograph is the Postmodern Sienkiewicz (*Sienkiewicz Ponowoczesny*) scholarly collection. It has its traditional form, available as open access in several formats, with all traditional elements of a scholarly publication; but the main outcome of the project was the digital collection.[36]

In both cases, the content consists of 12 dissertations on the writer's work and biography, and includes the same footnotes.

[35] 'The existing state of fragmentation sets limits on the retrieval, accessibility, and reusability of scientific resources and information, as well as constraining the possibilities for communication and cooperation between scholarly communities. The current diversity of formats, publication versions, content types, workflows, and operational models give rise to an increased necessity for the implementation of global standards which would set a common framework for scholarly communities operating in a digital environment.
Standardisation – perhaps the most significant, current challenge for digital humanities – requires a thoughtful, gradual course of action. It must be a process which includes a wide range of tasks, such as the identification of needs (that includes discipline-specific standards, already developed practices, and predefined formats); deliberation on, and adoption of, shared principles; and the implementation and promotion of newly introduced common practices and standards. Only such a pre-agreed general framework of rules, models, and practices can guarantee the reinforcement of openness and availability, and interoperability and interconnectivity of scholarly output and its further processability (content reuse, meta-search possibilities, long-term preservation)'. (Maciej et al. 2021a. 'Digital Scholarly Editions'. In *OPERAS-P Deliverable D6.5: Report on the Future of Scholarly Writing in SSH*, 158–170 ()).

[36] It was also analyzed in: Maryl, Maciej, Marta Błaszczyńska, Bartłomiej Szleszyński, and Tomasz Umerle. 'Dane badawcze w literaturoznawstwie' (Maryl et al. 2021b)

The biggest enrichment of this digital collection is the numerous visual materials (labelled with appropriate descriptions) – reproductions of paintings and objects, and photographs of the palace in Oblęgorek from the outside and the rooms used in the scholarly narrative. They are part of the scholarly narrative, but also may be treated as autonomous materials subject to possible further interpretation. Even though they are openly accessible, their interoperability and reusability is very limited; they can only be reused within the different collections on the same platform. This is due to the fact that the National Museum in Kielce, the owner of all visuals used in this project, for formal reasons, can not make them available for free, nor allow further sharing through the digital collection – thus, the rights to use them had to be purchased, which included a clause limiting their use to the NPLP.PL platform. In our work, we have also encountered attempts to include clauses in contracts for the use of visual materials which limit the duration of the contract (e.g. to 10 years); this type of practice, of course, not only prevents the reuse of this type of data, but actually does not allow for its use in long-term scholarly projects at all.

Another form of the Postmodern Sienkiewicz digital collection's enrichment, compared to that of a book, is the additional structuring. All articles have been written in a way which allows their individual fragments to be read also as autonomous texts which function outside the linear order of reading. Assigning them to categories permits a different order of reading based on thematic issues. This provides opportunities for some reuse of the collection's content – researchers can browse the texts for specific topics of interest in their research.

However, like all other collections on the NPLP.PL platform, Postmodern Sienkiewicz is open, all content is findable and accessible, but due to its main focus of telling a *digital*, *scholarly story*, there is much to be done in terms of interoperability and reusability. It seems that this is a broader problem in digital projects which were created and developed when reflecting on research data was not sufficiently developed.

Concerning lessons for the future, first of all, reflecting on data should be carried out as one of the research tasks when the entire project is being planned (retroactive FAIRification is often very difficult or simply impossible, and it is not easy to raise funds for it). Second, in the process of acquiring materials (visual, film, sound), it is worth trying to introduce clauses into the associated contracts which will allow their further sharing. Third, and finally, it must be accepted that full data FAIRness, while desirable, is rarely the primary goal of a research project – it will often have to be abandoned in order to achieve other, more important objectives.

A map-based digital collection

An interesting example of a digital collection (or a digital monograph) conceived of in a completely different way is the Atlas Literatury Zagłady (Atlas of Holocaust Literature). This is a digital born project, and it was only possible to create it in a digital environment, as it is based on a complex network of interconnections between entries on 'People' (osoby) 'Places' (miejsca), and 'Events' (wydarzenia). It belongs to the broadly defined discipline of literary studies, but could also be reassigned to historical studies or simply Holocaust studies. Its interdisciplinary nature is also emphasised by a map-based, topographical approach to the topics which have been studied and presented.

Entries under the 'Places' section consist of the following:

1. A static map with the location/place (concrete address, street, general area, for example 'Small Ghetto', and specific routes showing movement between destinations) marked on a map of the Ghetto;
2. An optional photograph of the place from the World War II period;
3. Excerpts from Holocaust testimonies in which the place appears. Each excerpt includes a) a description of the person who authored the testimonial from which the excerpt is taken, with a link to the entry about that person; and b) the period covered by the entry, with a link to an entry about that period;
4. Optional network of links to other places;
5. Bibliography.

Entries in the 'People' section consist of the following:

1. A person's bio (unstructured, prepared by scholars of Holocaust studies, based on multiple sources, and aimed at providing the most comprehensive set of information about a specific figure);
2. An interactive map which illustrates the places related to a given individual and which show a clear division between places inside the Ghetto and places on the Aryan outside, and shows places about which the individual wrote according to the period(s);
3. A network of links under a person's bio to places provides connections between an individual and specific locations in Warsaw they mentioned in their testimonies;

4) Bibliography.

Entries in the 'Events' section consist of a detailed presentation of historical events of that period and a network of links to places.

While the researcher-driven manual data preparation allows for the creation of a comprehensive and fact-checked bibliographical dataset (which has been verified by

a team of Holocaust scholars, digital collection experts, a graphic designer, and a cartographer), it presents several challenges from the point of view of FAIR principles. In this sense, it is rather a tool for researchers, and the data collected in it are reusable. It is also an open structure – it is easy to add new material (entries about places and people) on the basis of further elaborated Holocaust testimonies).

It seems that several actions could be considered to increase the quality of the data in this collection:

- Preparing and adding metadata to all maps;
- Adding DOI numbers to entries;
- Connecting entries with VIAF and Wikidata;
- Posting sets of entries to an open repository after the completion of each stage of work;
- Linking elements in the unstructured bios with other digital resources created in IBL PAN.

The Atlas of Holocaust Literature is also an example of a project which, after an initial pilot phase with relatively modest funding, lived to see its own continuation; it even expanded to more than ten times its original size. As there is already extensive material available after the completion of the pilot phase, it is possible to reflect in detail on FAIRification measures which could be implemented in phase two (which will last until 2027). A preliminary conclusion would therefore be that it is sometimes beneficial to conduct pilots for various projects, because, among many other aspects, they allow for better planning concerning data-related tasks in later activities.

3.1.4. Basing data classifications on real-life humanities research: a case of Polish literary studies

While the topic of data is being more and more often raised by humanists in Poland, who are being strongly encouraged to consider it by the funding and evaluation systems in which they operate (e.g. through regulations enforced by the Polish National Science Centre for their grantees – see Chapter 1), there have not been many attempts to organise the data-related knowledge for specific disciplines within the Humanities. Creating taxonomies or concrete classifications has not been a large part of data discussions till now. With rising data awareness has come a gradual shift however, and the narratives seem to become increasingly specialised and discipline-specific. We therefore present an exercise which we conducted to propose a data typology in Polish literary studies, which, acting as an example, may inspire other domains to undertake a similar endeavour.

3.1.4.1. Exploratory workshops – methodology

In autumn 2020, a team of members of the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN) organised a set of workshops for Polish literary scholars. Our aim was to propose a classification of data which would be rooted in the real life research topics provided by participants.

There were two main assumptions behind the project:

1. Polish literary scholars work with data, and the main types of such resources need to be identified for the community.
2. Literary scholars, themselves, are the best experts to help explore the diverse kinds of data they produce, analyse, or encounter in their research.

Following these assumptions, the call was open to all persons within the domain who were keen to participate in one or more workshops (there were three in total). The format was online, partly due to pandemic restrictions but also to encourage the involvement of scholars from different parts of the country. Participation in all three meetings was encouraged. The workshops included both plenary meetings with presentations, and breakout sessions for discussion. The discussions held in smaller groups were later summarised for the rest of the participants. During one of the exercises, scholars were encouraged to share specific examples of the data they worked with.

Unsurprisingly, the literary projects brought up as case studies were very diverse. Examples varied from semantics in Jane Austen's novels, through the analysis of religious metaphors, to studies of the Cairo music scene. Thus, the types of resources which the scholars collected, analysed, organised, and produced, and which they identified as 'literary data', differed greatly, including literary corpora with annotations (Jane Austen's novels), sources corpora with statistics on the use of different terms (religious metaphors), and lyrics and interviews (Cairo music scene), among others.

3.1.4.2. Our proposed taxonomy

It was the exercise above which allowed us to organise concrete examples of data which bore some similarity to each other into larger clusters. The proposed taxonomy, explored in detail in the article 'Dane badawcze w literaturoznawstwie' ('Research Data in Literary Studies') published in *Teksty Drugie* (Maryl et al. 2021b), includes six main types of data within literary studies: culture text, metadata, annotations, literary culture data, subject literature, and research documentation. These types are briefly presented below (please note that the descriptions and examples are not exhaustive).

Culture text – a concrete copy or edition of a text (text-based or based on graphics), for example, a published text, manuscript, theatre poster, 2D or 3D artefact, or audio or

audiovisual recording.

Metadata – data describing a specific copy/edition of a work or document, and/or technical details about its digital form.

Annotations – notes, comments, the critical apparatus, or the above mentioned TEI annotations which were added to the text during the interpretive or analytical stage, or automatically generated. They are mainly interpretation-based and rely on an accepted methodology rather than being based on a descriptive standard.

Literary culture data – sets of information about literary life such as calendars, lists of people and events, statistical data, survey results, or dictionaries of literary terms.

Subject literature – data which may include textbooks, scholarly commentaries, interpretations, and other studies of the subject of research.

Research documentation – workflow summaries, methods, notes, classification processes allowing one to learn more about the project and the way it was carried out, survey questions, meeting notes, or lists of participants.

3.1.4.3. Lessons learnt

We argue that the method applied to Polish literary scholars could be recommended to representatives of other disciplines who would also like to create a taxonomy for their field. Collaborative work based on real-life examples from researchers' lives allows taxonomies to be created which truly reflect the research reality and stresses the scholarly relevance of research data management, shifting it away from a purely administrative task and into the heart of the scholarly workflow where it belongs. A similar method was later applied in Italy by Gualandi, Pareschi, and Peroni (2022). Here the approach to the discussion about data, arts and humanities, and classification creation was an institutional, and not a disciplinary, one.

There is also a community-involvement aspect which has at least two advantages. First, scholars may participate in the creation of the typology and thus make it easier for the typology to be validated by the community. Second, the workshop, as a specific event, helps to raise awareness about research data management.

However, we have already noted some limitations to the developed typology. These might stem from the fact that only examples of research conducted by the participants themselves were discussed in depth. While there was a phase of reflection by IBL PAN organisers after the workshops, some examples of data which had not been raised at the workshop were probably overlooked. Another reason is that any kind of typology

creation exercise will, of course, encourage simplification. Therefore, the results of such exercises should not be treated as a finished product but rather as the beginning of a domain-specific data conversation which may be continued and developed by more case study examples brought up by the community.

3.2. Challenges in multilingualism

Erik Buelinckx, Francesco Gelati, Péter Király

Multilingualism is a central issue in the Arts and Humanities. Humanities research and data is very often written in national languages, and it will remain so for the foreseeable future. Moreover, most scholars deal with multilingual sources (see, for instance, Kuczycki et al. 2020). International online catalogues and databases are multilingual too. This multilingual obstacle in descriptive metadata can be solved by means of applying persistent identifiers to entities (such as people, place names, concepts). In this way, if the same metadata catalogue has different entries for the same item, say 'Léonard de Vinci' and 'Leonardo da Vinci', the use of persistent-identifier providers (internal or external, e.g., Wikidata and Geonames) will make both values findable. Linking catalogue entries to their controlled-vocabulary values is time-consuming and may require discipline-specific knowledge, but tools like OpenRefine and its reconciliation API make it easier.

Some controlled vocabularies, may they be interdisciplinary like Wikidata or discipline-specific like EHRI terms (Gelati 2019a; for EHRI see below), are multilingual; others, like the German Gemeinsame Normdatei, are monolingual. The DARIAH Thesaurus Maintenance Working Group's BackBone Thesaurus (BBT) focuses on identifying top-level concepts (facets and hierarchies), which will become a common basis for thesaurus building in an effort to meet the demands for objectivity and interdisciplinarity.[37]

In the arts, the multilingual Getty Arts and Architecture Thesaurus (AAT) is rather widely used, including by Europeana. Apart from English, the AAT has terms in Dutch, German, Polish, Spanish, French, and Chinese, among other languages, although, the percentage of translated items according to each language does change. Over the last two years, an initiative (BEINFRAT) has been created to bring together institutions from Belgium, Canada, Switzerland, and France who are working on, or interested in, the French translation of the AAT. The reason behind this was to make use of the existing knowledge in these (multilingual) countries, and to avoid work

[37] Christos Georgis, George Bruseker and Eleni Tsouloucha, 'BBTalk: An Online Service for Collaborative and Transparent Thesaurus Curation'. *ERCIM News*, 116, January 2019. Special theme: Transparency in Algorithmic Decision Making, ERCIM News 116, January 2019, URL Available Documents: BBTalk (ERCIM-News-116).pdf, BBTalk (ERCIM-News-116).docx

overlap. Ontologies are often published only in English, for example, the Shoah Vocabulary Specification for Holocaust studies and the Comic Book Ontology.

The most prominent example of a metadata catalogue in the Humanities is Europeana – an online platform where digital objects (and related metadata) relating to European cultural heritage are made accessible. Since 2021, Europeana has had an ongoing initiative called Europeana Translate, which 'aims to build connections between the Europeana and Automated Translation Digital Service Infrastructures (DSIs) to improve the usability of heritage resources by translating the metadata of the more than 25 million records available on Europeana'[38].

Europeana is built on linked data which supports multilingual data recording. Besides descriptive metadata, Europeana displays contextual metadata which links the topics, persons, places, and time spans which appear in the (free text) description to external vocabularies. The creation of contextual metadata can be supported by text mining methods, which finds similar entries in descriptive metadata and external vocabularies. Europeana prefers to use multilingual vocabularies, even if their expressiveness or other qualities are lower than the monolingual candidates, because they support the purposes of multilingual access, which is of more important value in this context. Europeana's practice is followed by other institutions, such as the Deutsche Digitale Bibliothek, which suggests that data providers in Germany use the same list of vocabularies. In a current metadata quality assessment project (in progress), records which make use of these multilingual vocabularies get higher scores than those which don't use them.

Vocabulary name	Base URL	Languages supported by the vocabulary
The Getty - Art & Architecture Thesaurus (AAT)	vocab.getty.edu/aat	Multilingual: English, Spanish, Dutch, German, Italian, French, Swedish, Chinese (written in Traditional and Simplified scripts etc.)
The Getty - Union List of Artist Names (ULAN)	vocab.getty.edu/ulan	Multilingual
Getty Thesaurus of Geographic Names (TGN)	vocab.getty.edu/tgn	Multilingual
Virtual International Authority File (VIAF)	viaf.org/viaf	Multilingual
Geonames	sws.geonames.org	Multilingual

[38] www.beeldengeluid.nl/en/knowledge/projects/europeana-translate

Vocabulary name	Base URL	Languages supported by the vocabulary
Iconclass	iconclass.org/	Multilingual
Gemeinsame Normdatei (GND)	d-nb.info/gnd	German
Israel Museum Jerusalem Concepts	museum.imj.org.il/imagine/thesaurus/	English, Hebrew
Library of Congress Subject Headings (LCSH)	id.loc.gov/authorities/subjects/	English
data.europeana.eu WWI Concepts from Library of Congress Subject Headings (LCSH)	data.europeana.eu	English, German, Dutch, Italian, French, Serbian, Danish
Europeana Sound Profiles	pro.europeana.eu/page/edm-profiles#sound-profiles	English, German, French, Spanish, Italian, Polish
UDC	udcdata.info	English
UNESCO Thesaurus	vocabularies.unesco.org/browser/thesaurus/en/	English, French, Spanish, Russian, Arabic
YSO - General Finnish ontology	finto.fi/yso/en/	English, Finish, Swedish
Wikidata	www.wikidata.org/entity/	Multilingual
Fashion Thesaurus	thesaurus.europeanafashion.eu/thesaurus/	English, Italian, Spanish, French, German, Dutch, Serbian, Swedish, Portuguese, Greek, Hebrew
Thesaurus of musical instrument names	vocabulary.mimo-international.com/InstrumentsKeywords/	English, Italian, French, German, Dutch, Swedish, Spanish, Catalan, Polish, Chinese

Within Europeana, there were several activities for improving the 'multilingual saturation' of the records. One of them has involved a proposal to use metrics in assessing multilinguality (see Király et al. 2019). The researchers suggested the following aspects for assessing individual metadata records: number of tagged literals (metadata field values having a language annotation), number of distinct language tags, number of tagged literals per language tag (synonyms), and average number of languages per property for which there is at least one language-tagged literal. The measurements showed, not surprisingly, that the contextual (entity related, subject indexing) metadata elements are more multilingual than the core content describing the elements (such as title, publication, extent); and that the result of the organisation's internal enhancement processes, which attempt to extract entities and inject terms from the multilingual

vocabularies back into the record (this part of the record is called Europeana proxy) are more multilingual, on average, than those parts submitted by the data providers. As a result, this process makes the record three to six times more multilingual in three of the four dimensions (the number of synonyms increased only by 25%).

The metadata schema used by library catalogues (MARC and PICA versions) are usually monolingual. For translations and multilingual publications it is possible to add different script variations (Cyrillic, Hebrew, Arabic, etc.), but there is no good, built-in solution for solving the general multilinguality problem. MARC however enables libraries to extend the schema with locally defined data elements. KBR (Royal Library of Belgium) created such data elements (a subfield $@, which could be applied to all MARC fields), where one can specify the language of the field which contains the ISO notation of the language (such as fr-BE or nl-BE). This is an example[39]:

110 $a Province du Brabant wallon $g Brabant wallon $c Wavre $@ fr-BE

110 denotes the main corporate name, one of those contextual entities which supports authority name control of the bibliographic records. Subfields a, g, and c stand for the name form, additional information, and location of the meeting, respectively. Unfortunately this approach does not have a long history, so, for the time being, the coverage of this language-denoting subfield is rather low.

The National Library of Israel (INL) has a somewhat similar approach.[40] They encode the writing script (instead of language) in authority controlled fields by using $9 (note: subfield 9 is left undefined in the MARC standard for all fields in order to make room for custom, library-defined information). There are four choices of script: Latin, Hebrew, Arabic, or Cyrillic (lat, heb, ara, cyr). Fields which are not authority controlled do not have the script encoded, but since the scripts are very different, when it comes to retrieving information based on script, criteria such as [a-z], [א-ת] etc. can be applied. This is an example[41]:

710 $9 ara $a مركز الديمقراطية وحقوق العاملين في فلسطين

710$a denotes the additional corporate name. The authority control coverage at INL is exceptionally high (97.94%), and the script is almost always encoded, even if there are multiple entities bound to a record[42], for example:

710 $9 heb $a לארשיב הוהי ידע תליהק

710 $9 lat $a Watchtower Bible and Tract Society of New York

[39] opac.kbr.be/LIBRARY/doc/SYRACUSE/20667684

[40] This section is based on email communication with Avaha Cohen, the Head of Cataloguing Section of the National Library of Israel.

[41] www.nli.org.il/en/books/NNL_ALEPH990000263960205171/NLI

[42] See for instance: www.nli.org.il/en/journals/NNL-Journals002880189/NLI

Of course recording the language or script does not necessarily mean multilinguality (in the above example, the two organisations' names are not translations of each other, but denote closely related, but distinct, entities, the Jehovah's Witnesses, and its American supporting organisation); however without it, it is not possible to move towards multilinguality.

A prominent example of a multilingual, archival meta-catalogue is the EHRI (European Holocaust Research Infrastructure) portal. Just as with Europeana, the EHRI portal collects and displays information from different data providers. And as with Europeana, data is provided in different languages. The EHRI portal's specificity lies in the use of the archival metadata standard XML-EAD (encoded archival description) in order to import and visualise data (Gelati 2019b). Multilingualism is, thus, primarily handled by mapping XML-EAD datasets, but multilingualism is treated differently in the EAD 2002 (EAD 2) and EAD 3 versions.

In EAD 2002 there are two different metadata fields concerning language. The field <langusage> (language usage) describes the language in which the archival item was described, whereas the field <langmaterial> (language of the material) enumerates the language(s) of the archival materials found in the unit being described. A British archival institution might use English (language usage) to describe a given archival collection which is made up of papers written in German and French (languages of the material). Avoiding confusion between these two metadata fields is the first challenge. Making sure that both fields are filled in, possibly with an ISO 639 language code, is another big challenge.

An EAD 2002 file can have no more than one language usage value. If parallel descriptions are required, i.e., to describe the same archival item in say French and Dutch (as is sometimes the case in multilingual Belgium), two distinct XML-EAD files need to be created.

In contrast, EAD 3 allows parallel descriptions to be embedded in the same XML-EAD file. Even though EAD 3 was released in 2015, the former version of the standard, EAD 2002, is still widely used throughout the archival community.

To circumvent the problem of (meta)data in multilingual environments not being good enough, artificial intelligence (AI) could be used in order to translate a search term (possibly making use of open multilingual thesauri and dictionaries). In fact, several research projects concerning archives, data, and AI are being conducted at present. See Colavizza et al. (2022) and European Commission. Directorate General for Communications Networks, Content and Technology. (2022).

3.3. Complexities in intellectual property and the application of regulatory frameworks in specific research scenarios – restrictions in text- and data-mining

Peter Gietz, Walter Scholger

3.3.1 Intellectual property rights (IPR) and thoughts on ownership

(Walter Scholger)

3.3.1.1. Introduction to IPR terms and legal systems

Researchers and students at universities and other educational and memory institutions are usually torn between advancing their own scholarly work and facing the complexities of reusing examples of other people's work which must be obtained and reflected upon in their research process. The dual role as **author** (or, more generally speaking, creator) and **user,** and finding a balance between these distinct interests, is the task of intellectual property rights (IPR) legislation, i.e., the definition of the rights of creators on the one hand and concrete exceptions to those rights, for example, for the research and education sector, on the other.

The *United Nations Universal Declaration of Human Rights of 10.12.1948*, reflects this tension in Article 27: 'Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits'[43]. This phrase is often cited in discussions about free access to educational and scientific resources, and as an underpinning of freedom of science, but it ignores the fact that the principle of 'free participation' is qualified in the following paragraph: 'Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author'.

While IPR covers a range of application areas – like trademarks and patents – this section will focus on **copyright**. When speaking of copyright, it is important to pay close attention to the accurate use of language: both in the media and in everyday conversations about this topic, people often talk about copyright and the subsequent use of copyrighted material, for example, for educational purposes, as being justified as **fair use**[44]. However, these terms are borrowed from Anglo-American law, which is based on a completely different legal system to that which is prevalent in Continental Europe. A more detailed explanation would go beyond the scope of this subchapter, but it should

[43] A/RES/217, UN-Doc. 217/A-(III), www.ohchr.org/en/human-rights/universal-declaration/translations/english
[44] For clarity, it should be mentioned that the term 'fair' in 'fair use' has no connotation to the FAIR principles acronym.

be noted that Anglo-American **common law**, as a utilitarian system, places competition and the interests of the market above the rights of the individual creator, whereas creators' rights are at the heart of the continental European legal system. This is evident even in the term 'copyright' (i.e. the 'right to copy'), whereas, Continental European **civil law legislation** focuses on the rights of authors/creators (Urheberrecht, droît d'auteur, szerzői jogok ...). Another difference is the rather broad definition of fair use (or fair dealing in Commonwealth nations) in exceptions to the legislation in the Anglo-American legal system; while Continental European legislation defines, very closely, individual scenarios which qualify as exceptions to and limitations on authors' rights, for example, for education and research purposes in the public interest.

3.3.1.2. Common ground

While the existence of distinctly different legal systems provides significant challenges in research and knowledge transfer scenarios, there is, fortunately, some globally valid common ground which is defined in international legal frameworks[45] and treaties, and supervised by the **World Intellectual Property Organisation** (WIPO)[46].

The term 'work' is used within the copyright context to refer to a wide range of intellectual creations, from literary works like novels to films, musical compositions, paintings, architecture, computer programs, and many more. At the heart of the definition of 'a work' is that it must be an **original intellectual creation**. Also, the legal protection extends only to **expressions**, not to procedures, mathematical concepts or raw data (which lack individual originality), the results of coincidental operations (lacking in intellectual intention), or ideas (lacking material creation). In our context, this will often mean that raw data is not subject to copyright protection, while curated data or data which is the result of a research process applied to this data – expressed, for example, in the form of a presentation, visualisation, or paper – is.

Metadata is an interesting and heavily debated subject in this context – while metadata which follows a strict metadata schema lacks the individuality required to afford it copyright protection, any metadata schema which allows for descriptive fields or individual allocation of designators – and hence, a sufficient degree of originality – may indeed be considered 'a work' in the sense of copyright legislation. This is even more intriguing because only a work can be **licensed**, while everything which does not constitute a work cannot. Current practice, however, tends to licence (even very restrictive and hence unoriginal) meta-data sets, which produces a certain tension between,

[45] The most prominent international treaties in this context are the Berne Convention for the protection of Literary and Artistic works, the WTO Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), and the WIPO Copyright Treaty (WCT).

[46] www.wipo.int/

on the one hand, making a commendable effort to signify that these meta-data sets can and should be re-used freely and without any restrictions, and, on the other, creating a grey area in attaching a licence to something which, under strict legal terms, cannot be licensed in the first place.

Another aspect of copyright is the rightsholder – in Continental European legal systems, the rightsholder is always the original creator of the work. The creator holds **personal rights** which are non-waivable (e.g. being considered the creator and naming the creation), and (largely economic) exploitation rights (e.g. reproduction, distribution, making available online), which can be transferred, for example, to an employer through an employment contract, publication contract (e.g. with publishing companies or collecting agencies), or licence agreements (with individual licensees for individual scenarios or exploitation methods). Since only the rightsholder can licence a work, regardless of how open or restrictive the licence, the role of the creator is vital to the publication and re-use of any work.

The duration of copyright is defined differently under different legislations; but in the vast majority of countries it is in effect until 70 years after the death of the original creator(s), or, in cases where either the original creator or their date of death are unknown, 70 years after the publication or creation of the work (keep in mind that the protection period lasts until the end of the calendar year of the 70th year). After that, works enter the public domain and can be used freely and, indeed, without acknowledging the original creators. While there may still be an ethical or scientific obligation to credit the original authors of **public domain** works – after all, it is common research practice to refer to one's sources, both to acknowledge the sources cited and for trustworthiness – there is no legal obligation to do so.

It is important to note that the protection periods apply to all original works, which is often a problem in the context of digitisation – while, for example, a mediaeval manuscript is, of course, public domain, its digitisation (i.e. scanning) may not be considered a mere reproduction, but an actual work (or effort, protected by comparable, so-called, neighbouring rights) in its own right. The reason for this is that the threshold for defining photographs as original creations is traditionally very low. There is a growing conviction among European Union legislators and scientific communities that any derivatives of public domain works should also be considered public domain, but (most) current copyright legislations do not reflect this – yet.

3.3.1.3. Exceptions for research and education

Another area of common ground provided by the aforementioned international treaties is the so-called three-steps test, which is used to determine exceptions to the rights of creators, for example, for research and education which is in the public interest. In short, countries may 'provide for limitations of, or exceptions to, the rights granted to authors of literary and artistic works under this Treaty in **certain special cases** which do not conflict **with a normal exploitation** of the work and **do not unreasonably prejudice the legitimate interests** of the author' (emphasis added).[47]

This is implemented under national copyright legislation through the definition of specific exceptions to copyright; the most well-known of these are free licences for personal use, personal use for research purposes, or for citing another person's works in the author's own scientific work – all of which are at the core of established research processes. Most copyright legislation also incorporates exceptions for the public re-use of protected works held in public libraries and educational institutions, as long as there is a strong **public interest** and there is fair **compensation** for the original rightsholder(s).

3.3.1.4. Legalising text and data mining

The European Union has introduced a number of **directives**[48] to provide some common ground across its member states (and the larger European Economic Area). While an actual common European legislation – as was achieved, for example, in the field of data protection through the introduction of the General Data Protection Regulation (GDPR), which will be elaborated on below – remains unrealised, these directives have indeed served to introduce common practices and terminologies, and impacted the way digital resources are created, published, and (re-)used in the context of research, education, and cultural heritage.

In the context of copyright, the most recent manifestation of the European Union Digital Agenda for Europe[49] is Directive 2019/790 of the European Parliament on copyright and related rights in the Digital Single Market[50], which, besides the vocal public discussions about upload filters, addressed a number of topics which impact directly on the field of Digital Humanities.

The most tangible innovations are calls for the EU member states to introduce **copyright exceptions for text and data mining** (Articles 3, 4) and for the **preservation of**

[47] WIPO Copyright Treaty (1996), wipolex.wipo.int/en/treaties/textdetails/12740

[48] EU Directive 96/9/EC on the legal protection of databases; EU Directive 2001/29/EG on the harmonisation of certain aspects of copyright and related rights in the information society; EU Directive 2003/98/EG on the re-use of public sector information; EU Directive 2012/28/EU on certain permitted uses of orphan works.

[49] www.europarl.europa.eu/factsheets/en/sheet/64/digital-agenda-for-europe

[50] EU Directive 2019/790 on copyright and related rights in the Digital Single Market

cultural heritage (Article 6). The former call addresses a growing need by our research community – especially our CLARIN-ERIC colleagues – for the automated and machine-based processing of texts (primarily), while the latter underscores the European Union's commitment to digitally preserve and, perhaps even more importantly, make accessible, cultural heritage resources.

The above directive allows for the following: 'analysing text and data in digital form in order to generate information which includes, but is not limited to, patterns, trends, and correlations'[51] by research organisations and cultural heritage institutions (note that the exception applies to institutions, not to individual researchers) if they have lawful access to the source (for example through university libraries' periodical subscriptions).

Perhaps even more spectacular, from a legal historical perspective, is the topic concerning the **use of works and other subject matter in** digital **and cross-border teaching activities** (Article 5), as it shakes the hitherto immovable condition that copyright ends at national borders.

3.3.1.5. Data ownership

Another related topic is the highly disputed subject of **data ownership**. While the rightsholder of any work is easily identified, data are not necessarily original intellectual creations. Even if there are no intellectual property rights attached to data sets, there is, without any doubt, a **value** to them (in this case, in the economic sense), which in turn can be exploited by their respective owners. There is currently no reliable legislation to resolve this issue, with the three main perspectives in the discussion claiming data ownership on behalf of a) the individual researcher(s) producing the data, b) the institution(s) hosting the researchers, or c) the funders enabling the research in the first place.

This discussion could, of course, remain an academic one if all the stakeholders (and possible owners) in the research process agreed on the best practice of **open data and open access**. In fact, the European Union expects its member states to '[make] publicly funded research data openly available, following the principle of "open by default" and compatible with the FAIR principles'[52], and calls for research data to be shared so that it is 'as open as possible, as closed as necessary'[53]. This strong commitment to openness in research practices is balanced by the responsible consideration of a legal area which is beyond, but at least as equally important as, IPR – the protection of personal data.

[51] ibid.
[52] EU Directive 2019/1024/EU on open data and the re-use of public sector information, Article 10
[53] ibid.

3.3.2. GDPR

(Peter Gietz, Walter Scholger)

3.3.2.1. Introduction

One of the most influential regulations in Europe, most prominently for the IT sector, is the EU General Data Protection Regulation (GDPR)[54], which has been in force since May 2018, and which has also had a big impact on research and education. It is one of the most comprehensive privacy preserving regulations in the world; its influence goes beyond European borders, as it also regulates every transfer of personal data between any EU country and any third country. Thus, basically, anyone outside the EU who has interactions with institutions within the EU must comply with the GDPR if any kind of personal identifiable information (PII) is exchanged. In any case, the GDPR has increased the awareness of data protection world-wide. Upon implementation, the regulation became applicable, binding law for all EU member states. Therefore, any research taking place within the EU also needs to comply with the GDPR[55].

3.3.2.2. The basic terms, 'personal data' and 'processing'

The regulation focuses on two basic terms: 'personal data' and 'processing'.

'Personal data' stands for 'any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'56; a natural person is defined as being alive, i.e., the personal data of deceased persons do not fall under the GDPR.

Personal data is not only personal attributes like names or contact data, but also, basically, any identifier which can be linked to such attributes. The IP address of a user can also be such an identifier, so that, for example, log files which contain IP addresses also need to be protected. There are different grades of sensitivity for personal data, and personal data defined as sensitive needs to be protected to an even higher degree and the processing is even more restricted, or actually prohibited. The GDPR considers personal data as sensitive if they reveal:

[54] Regulation (EU) 2016/679, see gdpr-info.eu/

[55] A good tutorial on GDPR and research by Walter Scholger and Sina Krottmaier can be found on a DARIAH-EU website: campus.dariah.eu/resource/posts/data-protection-in-research-practice-%E2%80%93-a-tutorial

[56] GDPR §4 (1)

- Ethnic origin
- political opinions
- religious or philosophical beliefs
- trade union membership

As well as

- genetic data
- biometric data for the purpose of uniquely identifying a natural person
- data concerning health
- data concerning a natural person's sex life or sexual orientation[57]

The lawful processing of such data is at the highest level of restriction, and usually requires the explicit consent of the data subject.

'Processing' stands for 'any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction'[58]; So basically everything which can possibly be done with personal data falls into the category of processing, and there is no difference whether this is done with digital data or with data written or printed on paper.

Other important terms are

- Data subject: the identified or identifiable natural person.
- Controller: 'a natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data'[59].
- Processor: 'a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller'[60].
- Consent (of the data subject): 'any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her'[61].

The GDPR restricts the processing of personal data by only allowing lawful, fair, and transparent processing. Every processing event needs to have a specified, explicit, and legitimate purpose and no processing beyond this is allowed. Luckily for research there is an exemption for archiving purposes which is in the public interest, scientific or

[57] GDPR §9 (1)
[58] GDPR §4 (2)
[59] GDPR §4 (7)
[60] GDPR §4 (8)
[61] GDPR §4 (11)

historical research purposes, or statistical purposes[62], which are all seen as lawful purposes. Besides this exemption, processing is only lawful if one of the following cases apply[63]:

- (a) with the consent of the data subject, which has to be freely given and informed;
- (b) in the case of a contract which was either signed or requested by the data subject;
- (c) where there has to be compliance with a legal obligation;
- (d) to protect the vital interests of the data subject or of another person;
- (e) to perform a task which is in the public interest;
- (f) in the case of 'legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject'[64] which 'shall not apply to processing carried out by public authorities in the performance of their tasks'[65].

As to the sensitive personal data mentioned above, the GDPR lists a much more restrictive set of exemptions[66].

Since the definition of personal data not only uses the term 'identified, but also 'identifiable', it is not only IP addresses which are defined as personal data, as stated above. Pseudonymised personal data, which are personal data which has been encrypted or otherwise modified in such a way that the link to a natural person is not obvious. But since pseudonymity can be reversed, for example, by an index linking a pseudonymous identifier with a clear name, it needs to be protected. This is not the case for completely anonymous data, i.e., data which have been irreversibly encrypted or otherwise modified so that attributes like the weight of a person can never be correlated with that person. If the person behind the data can not be identified, the data are considered anonymous and are not regulated by the GDPR.

3.3.2.3. GDPR key principles

There are seven key principles[67] in the GDPR (listed in §5) which provide guidance and clarify the intent of the GDPR but do not contain strict rules or instructions:

1. **Lawfulness, fairness, and transparency**:[68] which means that either the data subject must give their fully informed consent, i.e., understanding the reasons; or one of the other above-mentioned lawful processing categories must be met.

[62] See GDPR § 89
[63] See GDPR §6, numbering accordingly
[64] GDPR §6 (f)
[65] ibid.
[66] GDPR §9 (2) (a)–(j) lists these exemption from prohibition
[67] A good summary on these principles can be found at www.gdpreu.org/7-main-data-protection-principles-under-gdpr/
[68] As defined in GDPR §6

2. **Purpose limitation**: Personal data are only to be processed for a particular and legitimate purpose, and should not be processed beyond that purpose. An exception related to research and archiving is described in §89 (see below).
3. **Data minimisation**: Only those personal data really needed for a well defined purpose are to be processed. The data must, therefore, be limited to only the amount which is required for that purpose. The data must be adequate and relevant, but they also need to be sufficient for the purpose.
4. **Accuracy**: The data need to be accurate in the first place, regardless of the data subject's right to have inaccurate data corrected. If data are no longer accurate (e.g. an old postal address), they can, nevertheless, still be stored if they are needed for a specific purpose (e.g. as historical data).
5. **Storage limitation**: The data are not to be stored longer than the purpose allows for. Only three purposes allow for the indefinite storage of data (also see below about §89)
   a. archiving purposes in the public interest
   b. scientific or historical research purposes
   c. statistical purposes.
6. **Integrity and confidentiality**: The data need to be processed in a secure way so that they are accessible only to those with authorisation in respect to the data. This principle also means that data which is accidentally lost need to be recoverable, for example, via back-up mechanisms.
7. **Accountability**: All processors need to be instructed about GDPR, to take responsibility for their processing of personal data, and to follow well defined processes and measures which provide for conformity to the GDPR.

3.3.2.4. The rights of the data subject

Based upon the list of key principles in section 3.3.2.3, the GDPR outlines the following eight rights of data subjects:

1. **Right to be informed** (GDPR §15): The right to be informed about the processing of one's own personal data on request. Such a request must be answered, and if such data are processed, the answer needs to include the purpose, the kind of data, with whom the data is shared, how long the data are going to be stored, and the source of the data. It also needs to inform the subject about their rights concerning the rectification, deletion, or restriction of or objection to, the processing of their data; as well as their right to appeal to the regulating authority (see below). Last but not least, the answer has to state whether automatic decision making is being performed, together with information about the respective algorithms.

2. **Right of rectification** (GDPR § 16): The data subject is entitled to have incorrect data about them rectified. They can also demand to have missing data completed. This can be done through an application to the data processor.
3. **Right of deletion** (GDPR § 17): The data subject can demand to have their data deleted if the data processor has no legal basis for storing the data, if consent is being withdrawn, the retention period is over, etc. This is not the same as the right to be forgotten, which has also existed since 2014, when the EU Court of Justice ruled that old information which portrayed a person in a negative way should not be displayed by search engines, if requested by the data subject.
4. **Right to restriction of processing** (GDPR § 18): The data subject has the right to have the processing of their data restricted if the accuracy of the personal data is contested by the data subject, the processing is unlawful, there is no need for the processing according to the reasons provided by the processor, etc.
5. **Right to data portability** (GDPR § 20): The data subject has the right to receive a copy of their personal data, or have these transferred to another service provider. This is dependent on technical preconditions which might not exist.
6. **Right to object** (GDPR § 21): The data subject has the right to object to the processing of their data at any time, for example, processing for direct marketing purposes. If the processing is carried out for reasons of public interest, as, for example, in processing for scientific or historical research purposes (see below about § 89), such an objection will not hold.
7. **Right not to be subject to automated decision making** (GDPR § 22): The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, if it is not based on a contract or consent. This does not apply if such a decision making is authorised by the Union or Member States' law.
8. **Right to lodge a complaint** (GDPR § 77): Every data subject has the right to file a complaint with the supervisory authority in the Member State where they live, at their place of work, etc., if the data subject considers that the processing infringes the GDPR.

3.3.2.5. Research and GDPR, § 89

Taking into account both the strict principles of data processing and the rather extensive rights of data subjects, one may worry that any processing of personal data in research contexts carries the risk of infringing one or the other. Thankfully, in continuation of a recognisable trend in recent EU directives and regulations to facilitate research in the public interest, Article 89 of the GDPR calls for member states to introduce distinct exceptions in these contexts, while also ensuring that appropriate safeguards concerning data subjects are observed (e.g. data minimisation, pseudonymisation).

These exceptions are exclusive to 'archiving purposes in the public interest, scientific or historical research purposes or statistical purposes'[69] and are only intended to be applied when the previously mentioned rights of data subjects 'are likely to render impossible, or seriously impair' the processing purpose, for example, by rendering a study or data analysis invalid through the data subject withdrawing their consent to use their data in the research process.[70]

In addition to this definition concerning the specific processing situation which merits an exception, Article §6.1 (e), as seen above, broadly defines the situation where processing can also be considered legal if 'processing is necessary for the performance of a task carried out in the public interest'. While most research and education conducted at public research, educational, and cultural heritage institutions arguably qualify for this condition and, hence, satisfy the legal obligation towards the processing of personal data, it is generally accepted and expected practice among the Humanities and Social Sciences that research is conducted on the basis of the data subjects' informed consent in accordance with § 6.1(a). If a data subject withdraws their consent at a later time, the researchers who originally based their legally valid processing on given consent cannot then switch the legal basis of their processing to public interest in order to ignore their data subjects' wishes. It may therefore be preferable to base research scenarios on the member states' individual legislative implementation of § 89 rather than the general principles of § 6[71]. In any case, when conducting responsible research, it is essential to consider the ethical dimension of transparency and consent in addition to the legal context when processing personal data.

The ELDAH Consent Form Wizard is, on the one hand, a practical tool for fostering understanding about this delicate subject and, on the other, assists researchers in observing the legal obligations of the GDPR[72]. Developed as a cooperative project between humanities researchers, legal experts, and developers in the context of DARIAH--EU's working group on Ethics and Legality in Digital Arts and Humanities (ELDAH)[73], this tool supports humanities researchers within the European Union by guiding them through a questionnaire in order to create a valid consent form for processing personal data in the context of their specific professional activity which they can then provide to their data subjects.

[69] gdpr-info.eu/art-89-gdpr/
[70] For a more elaborated view, please consult the white paper authored by CLARIN Legal and Ethical Issues Committee members Pawel Kamocki, Erik Ketzan, and Julia Wildgans (Pawel Kamocki, Erik Ketzan, Julia Wildgans, 2018).
[71] The authors are aware that there is an ongoing discussion with very diverse viewpoints on this topic.
[72] consent.dariah.eu/
[73] eldah.hypotheses.org/

3.3.2.6. Data produced by the data subject

While the GDPR is mainly concerned with processed personal data, it does not make explicit statements about the ownership of the data. Thus, an as yet rather open question is whether the data subject owns the personal data relating to them or does it belong to the organisation which invested resources in collecting, refining, and analysing such data. However, 'more and more legal experts argue that personal data is owned by the data subjects, rather than the data controller'[74]. This is also relevant to the data collected by mobile apps, for example, data on whether a particular place is less or more crowded than usual. Such data have been measured and gathered by the users of mobile devices which are usually owned by that user. It seems unfair that these data are in the hands of very few companies, which can then exploit them according to their commercial needs. Should it rather not be the public, and with that, also researchers who have the right to own such data so that they can access them, without any filtering or advertising, as open access data? It should be possible for individuals to donate to public platforms the data measured by their own devices (of course in an anonymous form). Such data will certainly also be relevant for research in the Humanities and Social Sciences.

The question of ownership of data is yet to be answered, and it is a complicated question. Various roles, for example, creating, using, compiling, funding creation or compilation, or being the subject of the data, could claim ownership, and such claims could include unforeseen risks, such as liability, as discussed by Kevin M. Alvero[75]. This discussion is also highly relevant with respect to Internet of Things[76].

3.3.3. Access control to research data in the frame of FAIR principles and open access

(Peter Gietz)

Research and, hence, the creation of research data happens within different eco systems, which have different approaches when it comes to authentication (checking and proving identity) and authorisation (checking and enforcing the identity's access rights to resources). The following describes two such ecosystems, namely research infrastructures and open research, and then proposes the use of common technologies

[74] Egil Bergenlind, *Who owns your personal data under GDPR*; July 3, 2017, www.dporganizer.com/blog/gdpr/who-owns-personal-data/ (Bergenlind 2017)

[75] Alvero, Kevin M., *Data Ownership: Considerations for Risk Management*, ISACA Journal, 2020,2, 1. April 2020, www.isaca.org/-/media/files/isacadp/project/isaca/articles/journal/2020/volume-2/data-ownership_jo a_eng_0320.pdf

[76] Thomas J. Farkas. 2017. *Data created by the internet of things: the new gold without ownership*, Revista La Propiedad Inmaterial, August, revistas.uexternado.edu.co/index.php/propin/article/view/4975

when it comes to authenticating authors, data creators, and commentators of texts and data. As FAIR principles have become common ground in arts and humanities research, it has also formed the framework for this subchapter, which describes the necessity for authentication and authorisation, even within the frame of open access. Basically, the technologies developed within the framework of research infrastructures, especially so called federated identity management (FIM), are useful at different stages of the production of research data.

3.3.3.1. Ecosystem research infrastructures

Starting in the mid twentieth century, a considerable shift in research infrastructure has taken place in general, which can also be observed in the Arts and Humanities – earlier, scholarly communication had been print-based only, but researchers in the nineteen fifties started to digitise data so they could be processed by computers, and thus paved the way for Digital Humanities[77]. While the first such humanities researchers used mainframe computers, the later evolution of the Internet changed all our communication habits. When the Internet became accessible to humanities researchers, a second shift towards more collaborative research practices took place. Both these shifts (print to digital, and researcher in the attic to researchers collaborating remotely via Internet communication tools) were accompanied by the evolution of humanities research infrastructure, which provided, firstly, the tools to manipulate digitised data and secondly collaboration platforms. Such issues will become even more prevalent in the context of the nascent European Collaborative Cloud for Cultural Heritage (ECCCH)[78]. In many fields of research, not only have virtual research environments (VRE) been established, but also virtual research infrastructures (VRI), which provide general services for the VREs. With VRIs came another mental shift in research, initiated by grid computing, which is now ubiquitous in the form of cloud computing; along with this, in many instances the researcher cannot tell for sure anymore where exactly the research and cultural heritage data are physically stored, which is an issue with regard to trust and data privacy. Thus it is very relevant who provides cloud resources; and the current market situation is that, as of 2021, three US companies (Amazon, Google, and Microsoft) control 63% of the market[79], and since even more US companies have additional

[77] Father Roberto Busa's *Index Thomisticus* is seen as the earliest use of computers for processing natural language and literature, and thus as the foundation of Digital Humanities, see (Busa 1980). It also laid 'the groundwork for a profound epistemological and cultural transformation' as argued by, for example, Puthiya Purayil Sneha, 'The Digital Humanities from Father Busa to Edward Snowden', *Media Development*, Vol. LXIV 2/2017. Published on May 13, 2017, waccglobal.org/the-digital-humanities-from-father-busa-to-edward-snowden/. A comprehensive and critical evaluation of devalued, feminised labour which was contributed to Busa's *Index Thomisticus* can be found in Nyhan, Julianne: Hidden and *Devalued Feminized Labour in the Digital Humanities - On the Index Thomisticus Project 1954–67*. London, Routledge, 2022.

[78] See ec.europa.eu/commission/presscorner/detail/en/IP_22_3855

[79] See www.datacenterdynamics.com/en/news/amazon-microsoft-google-dominate-cloud-market-post-str ong-results/

shares in this market, the situation must be seen as highly problematic. In such a market, it is vital that research data are stored at more trustworthy and more GDPR--compliant providers, such as European research computing centres which provide cloud services, for example, within the framework of the EOSC (eosc-portal.eu), the European Open Science Cloud.

All this took place across every field of scholarly research, in the natural sciences as well as in the Arts and Humanities, so that in addition to the more domain-related VRIs like those set up by DARIAH-EU and CLARIN-ERIC, VRIs were also developed which provide basic generic services to all fields of research: EGI (www.egi.eu) for advanced computing resources, EUDAT (eudat.eu) for storage resources and data sharing, and OpenAIRE (www.openaire.eu) offering services for open research. These three are now combined within the EOSC framework, an environment for hosting and processing research data to support European research. While EOSC wants to 'enable a trusted, virtual, federated environment in Europe to store, share and reuse digital outputs from research (including publications, data, metadata and software) across borders and scientific disciplines'[80], dedicated domain specific VRIs have also been established in almost every field of research. Here we can distinguish between general services (computing resources, batch run infrastructure, [scheduler], storage resources, research data annotation, data replication, long term preservation, persistent identifiers, cloud storage, data repositories, data management planning, data discovery, identity management and authorisation, etc.) and domain specific services (domain specific metadata standards, collaboration platforms, research software, repositories, authorisation, etc.).

Most VRIs have a number of features in common, a prominent one being the integration of authentication and authorisation infrastructure. Within the European project AARC (Authentication and Authorisation for Research Collaborations, aarc-project.eu)[81], a blueprint architecture has been established which allows for interoperable authentication and authorisation infrastructures (AAI) based on federated identity management technologies such as Security Assertion Markup Language (SAML)[82] and OpenID Connect (OIDC)[83]. A good overview of the AAI activities within domain specific VRIs can be found in FIM4R's activity, where a number of VRIs have come together in workshops to align their individual AAIs and formulate their specific requirements for federated identity management[84].

[80] See www.eoscsecretariat.eu/sites/default/files/open_consultation_booklet_sria-eosc_20-july-2020.pdf, p. 4
[81] Actually two funding phases, AARC I 2015–2017 and AARC II 2017–2019.
[82] See wiki.oasis-open.org/security/FrontPage
[83] See openid.net/developers/specs/
[84] See fim4r.org/

3.3.3.2. Ecosystem open science, open research

Open science – or open research, to better represent the Arts and Humanities – is part of a broader movement comprising many more aspects of society. This movement has common features, or sub aspects, such as[85]:

- Open source
- Open standards
- Open access
- Open data
- Open content
- Open knowledge
- Open education
- Open innovation

For the following, the most relevant aspects are open standards and open source, together with the observation of the intellectual property and regulatory frameworks as described in 3.3.1, especially the privacy legislation specified in the General Data Protection Regulation (GDPR).

Any interoperability between different AAIs is dependent on the support of open standards, in this case mainly SAML (Security Assertion Markup Language)[86], as it is the most established protocol in the realm of research and higher education. Open ID Connect (OIDC)[87] and the System for Cross-Domain Identity Management (SCIM)[88] are evolving standards which will play an increasing role in future AAIs.

Open source has always been elementary to research environments, for good reason:

- transparency
- flexibility
- independence
- intersubjectivity
- sustainability
- promotion of innovation
- affordability

[85] A more complete list of everything connected with open science can be found in the output of the EU's project FOSTER (Fostering the Practical Implementation of Open Science in Horizon 2020 and Beyond) at www.fosteropenscience.eu/foster-taxonomy/open-science-definition.

[86] See saml.xml.org/saml-specifications

[87] See openid.net/developers/specs/

[88] See www.simplecloud.info/

3.3.3.3. Authentication and authorisation in an 'open' world

As has been mentioned in this document several times, FAIR principles have become common ground for RDM. The 'A' in FAIR stands for 'Accessible', which has the following aspects:

- a retrieval interface
- a standardised communication protocol
- authentication and authorisation
- the long term preservation of metadata

Open access is defined as 'a comprehensive source of human knowledge and cultural heritage which has been approved by the scientific community'89. Arguably, any publicly funded information – i.e., by means of taxpayers' money – should in turn be openly accessible to the taxpayer (in a global sense). This also leads to conflicts of interest between libraries and scholarly publishers, and results in new business models for the latter.

If information should be open for all, why would you need authentication and authorisation (A&A) then? Even with open access platforms, access control makes sense for several reasons:

- Proof of authorship-
  Content being uploaded to an open access platform anonymously would mean that its authorship could not be proven. Thus, at least the content upload needs to be secured with A&A.
- De-anonymise scholarly discussions-
  As soon as an open access platform allows content to be commented on, it is, again, important for everyone to know who takes part in the discussion, just as it was in the times when reviews were made in printed journals.
- Secure pre-publication content-
  Many open access platforms have the possibility of enforcing access restrictions on yet to be published content, so that, for example, content can be shared with some colleagues ahead of the official publication.
- Payments-
  Part of the business model of commercial open access platforms is that the author needs to pay for the publication of content, which again requires A&A.
- Non-open materials-
  And of course, some materials cannot be openly accessed because of the sensitivity of their data, because of restricted copyright, or because they are embargoed or contain delayed access materials.

[89] See Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, openaccess.mpg.de/Berlin-Declaration

In those areas where A&A is needed, it has to be done right, i.e., using standards and being interoperable with other research infrastructures while reusing existing infrastructures. Following the AARC Blueprint Architecture allows for maximum interoperability.

3.3.3.4. The technical basics of AAI

Identity management (IdM) is a system for managing computer records which map the identities of people. These records contain unique names (e.g. login name), credentials (e.g. password) and any other necessary information about the person, such as name, email, group memberships, etc. IdM takes place within an organisation.

Federated identity management (FIM) is a system which allows identities to be used beyond organisational borders – the users authenticate themselves by proving their identity to a technical system, the so called identity provider (IdP), which is connected to the IdM of the user's home organisation. This allows them to access the services of other organisations, which decide to allow or restrict access based on the authorisation information from that IdP. The communication between the IdP and the actual service is intermediated by a service provider (SP) which supports the respective communication protocols. A nice side-effect of FIM is the single sign-on (SSO) function, because once a user has been authenticated at an IdP, the IdP will store this information for a reasonable amount of time and provide it to any other SP the user wants to access without requiring the user to authenticate again.

A structure which allows for FIM is called the authentication and authorisation infrastructure (AAI) and is a collection of IdPs and SPs which build a federation based on trust through contracts: the IdP states that the information provided to the service is accurate and current, while the SP states that they will use the data provided by the IdP only for the agreed use case, especially any personal identifiable information (PII), which is governed by the GDPR. Besides IdPs and SPs, a federation needs central management components to manage contracts (1 to n instead of n to n), maintain the federation membership list, manage the technical data about the computer systems involved (URLs, server certificates), and operate a central discovery service (DS), through which the user can select their home organisation. A federation is thus a group of organisations running IdPs and SPs which agree on a common set of rules and standards.

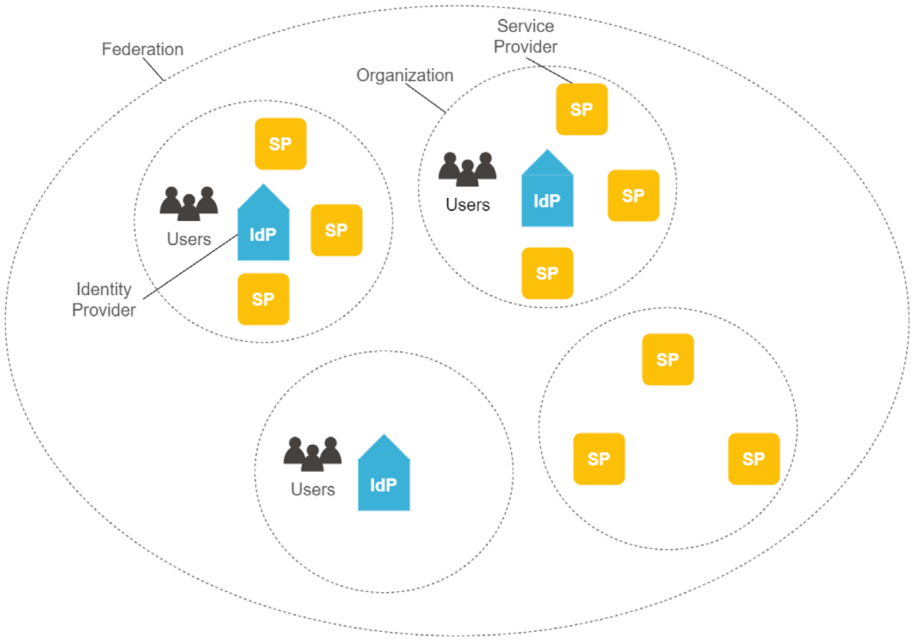


Figure 4. DAASI International, *Federated Identity Management*.

Almost every national research network operates such federations and all these federations are connected via the inter-federation eduGAIN[90]. With eduGAIN we have a world-wide infrastructure which can be used by any application in research and higher education, thus, also, by virtual research infrastructures and open access platforms.

3.3.3.5. The AARC project, blueprint architecture and future trends in AAI

Based on requirements postulated by the FIM4R community[91] and based on the findings of an EU[92] funded study on this matter, which was performed by national research networks and eduGAIN to facilitate VRIs, the EU funded two phases of the AARC project[93], where, during the first phase, the requirements for federated authentication and authorisation were gathered from e-infrastructures. Twenty partners joined the project, including the National Research and Education Network (NREN); GÉANT; e-infrastructures such as EGI, PRACE (prace-ri.eu), and EUDAT; and important user communities like ELIXIR (elixir-europe.org) and DARIAH-EU.[94]

[90] See edugain.org/
[91] See 'Federated Identity Management for Research Collaborations', 28 August 2013, fim4r.org/wp-content/uploads/2017/07/CERN-OPEN-2012-006-2.pdf
[92] See European Union, 'Advancing Technologies and Federating Communities - A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe', 2012, wiki.geant.org/download/attachments/21266435/2012-AAA-Study-report-final.pdf?version=1&modificationDate=1355507360046&api=v2
[93] AARC1 May 2015–April 2017, AARC 2 May 2017–April
[94] See aarc-project.eu/about/why-and-how/

Based on these requirements, a blueprint architecture (BPA)[95] was created. It defines the functional building blocks for VRIs so they can achieve maximum interoperability with the national AAIs and with eduGAIN, arranged in key components which can be used according to the specific requirements. The general architecture includes so--called proxies, which allow for very flexible architectures. Proxies can hide technical complexities from federation IdPs and SPs while communicating with standard protocols; they can also support different protocols, so that, for example, applications only supporting OIDC and applications supporting only SAML can be included in the user's SSO experience.

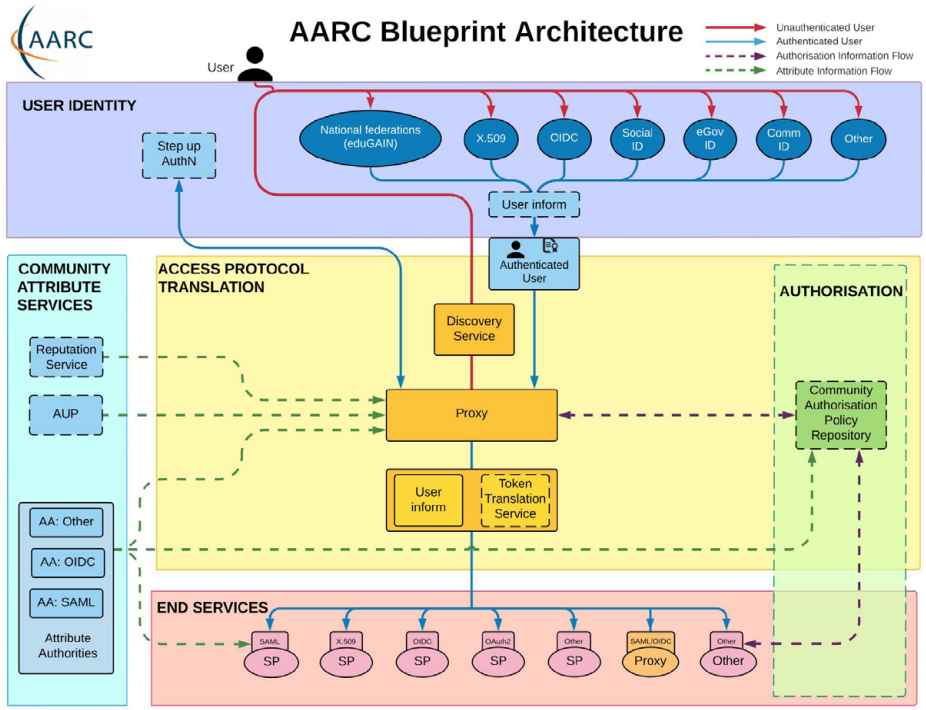


Figure 5. AARC Community members, & AppInt members. (2019). *AARC Blueprint Architecture 2019 (AARC-G045)*. DOI: *10.5281/zenodo.3672785*

The BPA consist of five component levels:

- User identity - different protocols for IdPs and other authentication systems.
- Community attribute services - additional attributes which add to the attributes already stored in the home organisation, which is especially useful in regard to virtual organisations (VO), consisting of members of various real life organisations, such as multilateral research projects. This allows for privilege group management in VOs.
- Access protocol translation for multi-protocol support.
- Authorisation allowing control access to services and resources.
- End services, the actual SP protected applications.

[95] See aarc-project.eu/architecture/

Most modern VRIs, including the new generic EOSC AAI and also the domain specific DARIAH AAI, already support the BPA proxy model, which allows for interoperability. This means that, for example, DARIAH-EU users can access EGI services via their home organisation account. Other users can access DARIAH services or the SSH Open Marketplace[96] via the same technology. Another important result of AARC is the policy toolkit which provides concise and correct templates for different policies, reflecting the best, common practices in federated contexts. Such common policies can harmonise the reliability of identities and attributes but also other aspects of trust in federations. The AARC results had a significant impact, and community AAIs should adhere to these best practices.

Other future trends in AAI will be the introduction of life-long identities via eduID systems, which will allow users to keep just one identifier when switching from one research organisation to another, thus keeping more control of their personal data. This type of eduID Idp can also function as a proxy and is thus compatible with the BPA. New protocols will also play a greater role in the future – OIDC for SSO systems beyond Web-based applications, and SCIM for more reliable and scalable deprovisioning of user entries in applications.

3.4. Incentives (or the lack thereof) to publish research data, as, traditionally, they have no 'real place' in Humanities' scientific communication

Marta Błaszczyńska, Vera Chiquet, Erzsébet Tóth-Czifra, Ulrike Wuttke

Clearly, the limitations of current research assessment and rewards criteria have been recognised as the Achilles heel of the process of firmly grounding data sharing and other open research practices in research realities. In academia, scholars are still facing conflicting expectations. On the one hand, we see advocacy and research policy efforts[97] repeatedly trying to reform the current system, which is dominated by publisher prestige, to, instead, reward scholars for all activities and content types involved in research processes, not only for publications in the traditional, print legacy sense. These attempts have given rise to initiatives such as the Open Science Career Assessment Matrix (OS-CAM), but, as pointed out by Schöpfel and Azeroual (2021), such attempts do not cover the detailed and domain-specific guidance which could

[96] See marketplace.sshopencloud.eu/

[97] Most notably, the San Francisco Declaration on Research Assessment, the Leiden Manifesto, and Metrics Tide. From a Social Science and Humanities perspective, see Emmanuel Kulczyczki's works, such as dariahopen.hypotheses.org/1582.

make them truly operational – for instance, exactly which people and which data-sharing behaviours should be rewarded (academics in the strict sense, or including research support personnel) – nor do they address anomalies such as the potential dis- tortion of research interests in easily accessible and shareable data sources (Edmond 2015) as an entailment of data sharing mandates. Yet, the biggest problem is that although we see investment in data management and sharing increasingly become a condition of external research grant funding, they largely remain invisible when it comes to academic institutional hiring, tenure, and promotion criteria. The Coalition for Advancing Research Assessment (CoARA), established in 2022, and the associated Europe-wide reform of research assessment, has the potential to become a game--changer in this respect.

In this chapter, we highlight some examples of efforts and initiatives which aim to incentivise and reward data sharing, encompassing the professional contexts of the Working Group's members. Although these are not directly linked to any institutional or research funder's assessment criteria, they can inform and facilitate the ongoing reforms from a domain-specific point of view.

3.4.1. A non-exhaustive panorama of data journals in the Arts and Humanities

Probably the most well-known instrument to gain recognition for data sharing is to 'gift wrap' data sets as data papers in data journals and thus align them with the well--established scholarly journal format and all its information management entailments (discovery, indexing, and citation tracking systems which are optimised only for papers and are included in research tools). In this way, they enter the scholarly citation system, which is still an absolute necessity if they are to receive proper academic credit.

Below is a sample of data journals with an arts and humanities scope:

Research Data Journal for the Humanities and Social Sciences
Journal of Open Humanities Data Dataverse
Journal of Cultural Analytics
Journal of Open Archaeology Data Journal of the Text Encoding Initiative
(Video) Journal of Embodied Research
RIDE – A review journal for digital editions and resources
Data publications in the Zeitschrift für digitale Geisteswissenschaften

3.4.2. LexSeal – an evaluation framework for the assessment of lexicographical datasets

As mentioned in the introduction to this chapter, an important dimension of data sharing rewards which is, as yet, largely missing, is to have detailed discipline specific or data type specific evaluation frameworks in place. The ELEXIS Lexicographic Data Seal of Compliance is a response to this need. It puts forward an evaluation framework for the quality assessment of lexicographical datasets and creates a community-based certificate of compliance using the best scholarly practices to be awarded to individual lexicographic datasets in recognition of their creators' self-assessed and well-documented adherence to the principles of trustworthiness, interoperability, stewardship, citability, reciprocity, and openness. You can read more about this framework in Tasovac et al. (2021).

3.4.3. OBERRED – Open Badge Ecosystem for the Recognition of Skills in Research Data management and sharing

A way of bringing open badges (OBs) into the world of research data management, although not just in an arts and humanities context, has been provided by the OBERRED (Open Badge Ecosystem for the Recognition of Skills in Research Data Management and Sharing) project. The project was run between 2019 and 2022, financed as part of the Erasmus+ programme, and coordinated by the Côte d'Azur University. OBERRED aimed '**to create an open badge ecosystem for the recognition of skills in sharing research data**'.

OBERRED lists a number of advantages to using open badges in research data management. Apart from the benefits which OBs have under any circumstances (**scalability, openness, interoperability, and the aspect of having something visual to display**), they serve the specific context of research data management very well. RDM competencies are often still not widely standardised or recognised (and sometimes even remain marginalised) within the whole academic community. Open badges allow for a system which, on the one side, provides a structure for competence recognition and, on the other, remains flexible enough to allow for national, institutional, and role-based modifications (e.g. an open badge about data storage may be adjusted to include information about the organisation or the country's repository). They also help to organise the different competencies associated with data curation, enrichment, and publication (as has been done within OBERRED's skills framework).

An RDM-related skills framework has been developed in OBERRED based on the data-cycle. The identified competences allowed the team to develop an ecosystem consisting of 22 badges acknowledging concrete skills related to data planning and designing, collecting and managing, describing, formatting and storing, quality assurance, processing and analysing, archving, publishing, and ensuring the discoverability of data. The team also created three massive open online courses (MOOCs): Open Badges for Open Science, Basics of Managing and Sharing Research Data, and Being an Animator of the Ecosystem. Plans for further implementation are being discussed.

3.4.4. Participatory knowledge practices in analogue and digital image archives

Another example for incentivising data sharing is to provide open archives, where research data can be shown in new ways. That is what the SNSF project called Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA, about.participatory-archives.ch, work in progress) is trying to implement. The PIA project connects the world of data and things in an interdisciplinary manner. First it explores the phases of the analogue and digital archive from the perspectives of cultural anthropology, technology, and design. By looking at the participatory knowledge practices in image archives in an interdisciplinary way, it engages with the processes of unfamiliar disciplines and strives to cooperatively implement them. For this purpose, digital tools are being developed which support the contextualising, linking, and contrasting of images. It is dedicated to bringing about collaboration between the scientific community and the wider public, facilitating the preservation and dissemination of knowledge, and encouraging users to engage collaboratively with their own history and contemporary practices. In a series of workshops and interviews with future users, the new requirements of digital and process-oriented knowledge production will be elaborated.

Using three collections as examples, the PIA develops interfaces which enable the collaborative indexing and use of archival materials. The interfaces – the graphical user interface and the application programming interfaces (APIs) – provide tools and visual interfaces for the collaborative production and visualisation of knowledge with the aim of enabling a reflective and intuitive experience. Like all archives, this example holds a wealth of metadata which can be analysed. The development of a participatory archive platform such as the one being carried out by the PIA research project, requires a flexible infrastructure which allows genuine data curation and a robust underlying data model. Therefore, it is critical to use linked open usable data (LOUD) standards, such as IIIF, Linked Art, or the Web Annotation Data Model, which help in the dissemination and reuse of cultural heritage resources, as well as contributing to the

sustainability of digital humanities initiatives (Raemy 2021).

What remains unaddressed, however, is the wealth of (historical) knowledge contained in this data, which can be continuously enriched, reflected upon, and contextualised. Therefore, digital means are being developed which allow different stakeholders to freely access and interact with the project's data. Both humans and machines can use, contribute to, correct, and annotate the existing data in an open and interoperable manner, thus encouraging exchange and the creation of new knowledge. To do this, web-based standards are being used which have already been widely adopted in the cultural heritage field. Currently, the project uses Omeka (omeka.org), an open-source web publishing platform for sharing digital collections and creating media-rich online exhibits: explore.participatory-archives.ch/s/explore.

3.5. Good practices in data co-curation between cultural heritage institutions and researchers

Erik Buelinckx, Vera Chiquet, Rita Gautschy, Erzsébet Tóth-Czifra

Cultural heritage institutions (CHIs) are homes for cultural artefacts. CHI collections can include artefacts from different disciplinary fields, like art, photography, visual or performing arts (dance, theatre, cinema), or design, fashion, music and so on. For the dominant part of research fields belonging to the Arts and Humanities domain, such cultural heritage collections are research data (Tasovac, Chambers, and Tóth-Czifra 2020). The digital availability of these materials along with clear reuse conditions is an essential precondition of FAIR and open data workflows in these disciplines.

The biggest obstacles to the productive reuse of digitised cultural heritage resources, from which many others derive, are the legal and ethical restrictions in which the use conditions of cultural heritage sources are embedded. Determining ownership status over research which is based upon such material often poses many challenges, as ownership is, on some level, shared between the researcher who carries out the scientific analysis on the source materials, the institution which hosts and curates this material, and the people and cultures who gave rise to the objects in question (e.g. photographers and, also, the subjects of the photographs). A great deal of ambiguity surrounding ownership, e.g. the identity of copyright holders, licensing, and other possible complexities under reuse conditions, defines all the further steps in research and data workflows, along with possibilities of facilitating access or even publishing scholarship concerning these cultural artefacts. As such, this can be recognised as being a recurrent problem for all stakeholders working with cultural heritage data.

In simple cases, the rights must be obtained, for example, from the creator and from the institution which owns the object in order to move towards digital openness. Documentary photographs of a vernissage or a film premiere, however, may constitute much more complicated cases – what about the personal rights of the persons depicted; is there also a work of art photographed in the background; may this be reproduced; who was the photographer; on whose behalf were the pictures taken; was a declaration of consent signed; who owns these pictures?

So far, the decision to open up collections for digital reuse has been rather like the Wild West; certain institutions are very open, like the Rijksmuseum (www.rijksmuseum.nl), while others remain resistant or simply lack the capacity. Public cultural heritage institutions are asked to follow EU directives and their national implementation (see above). The main problem starts to become increasingly less that these public CHIs are unwilling to make the data open (based on notions like 'MY data'), but more often the impossibility of doing so due to a lack of funding (financial, personnel, etc.)[98]. In Belgium, for instance, discussions are ongoing for obtaining additional funding for opening up data based on the experiences of the meteorological institute. The institute was partly funded through the selling of data, but, as an alternative, they calculated what was needed to continue their work following an alternative business model which is not based on selling data but instead making data available openly. To explore the possibilities for implementing such models, they created a fairly simple table referring to personnel (data scientists, IT people, documentalists, etc.) and material (servers, data storage, etc.), depending on the amount of open data which needs to be made available. Besides, in relation to the policy environment, a conflict between Directive 2003/98/EC on the re-use of public sector information (known as the PSI directive) and the REC 2011/711/EU calling for digitisation efforts, adds further complications, and even challenges, to making cultural heritage data openly available as research data. The former is still not very restrictive concerning museums, archives, and libraries which charge higher fees for the re-use of their holdings, and allows them to conclude exclusive agreements for digitising their material. Evidently, charging for reuse permissions as an existing business model for cultural heritage institutions will remain in sharp conflict with supporting open and FAIR research mandates and other public digitisation efforts.

Progressing in the direction of opening up cultural heritage collections and investing in the clarification of reuse rights, the Orphan Works Directive (en.wikipedia.org/wiki/Orphan_Works_Directive) is an important tool, even though its actual implementation

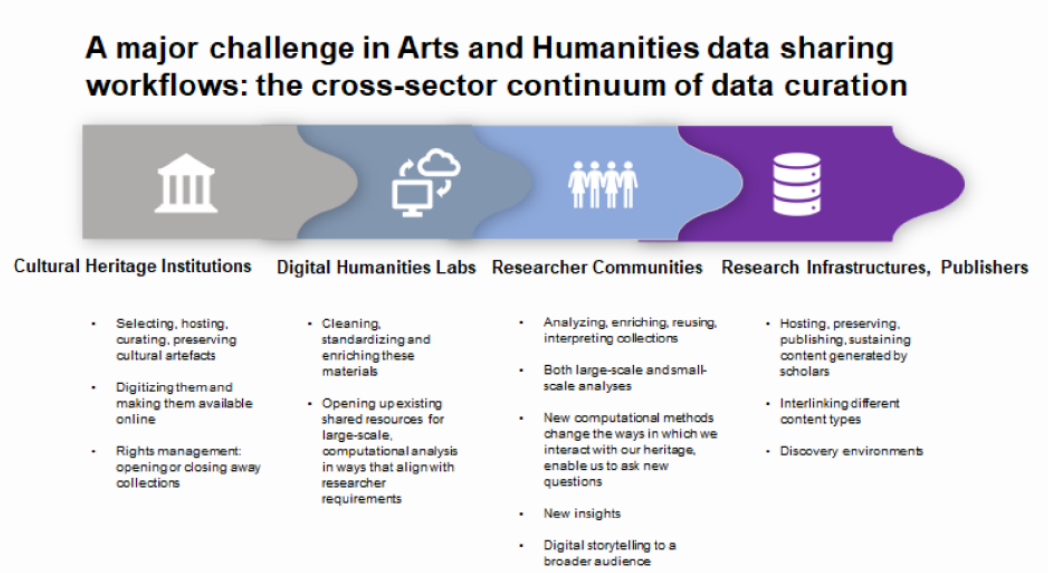
[98] See, for instance, a recent example: Stakeholders' Survey on a European Collaborative Cloud for Cultural Heritage: Report on the online survey results (op.europa.eu/en/publication-detail/-/publication/06851eec-7f4d-11ed-9887-01aa75ed71a1) but see also Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, et al. 2015.

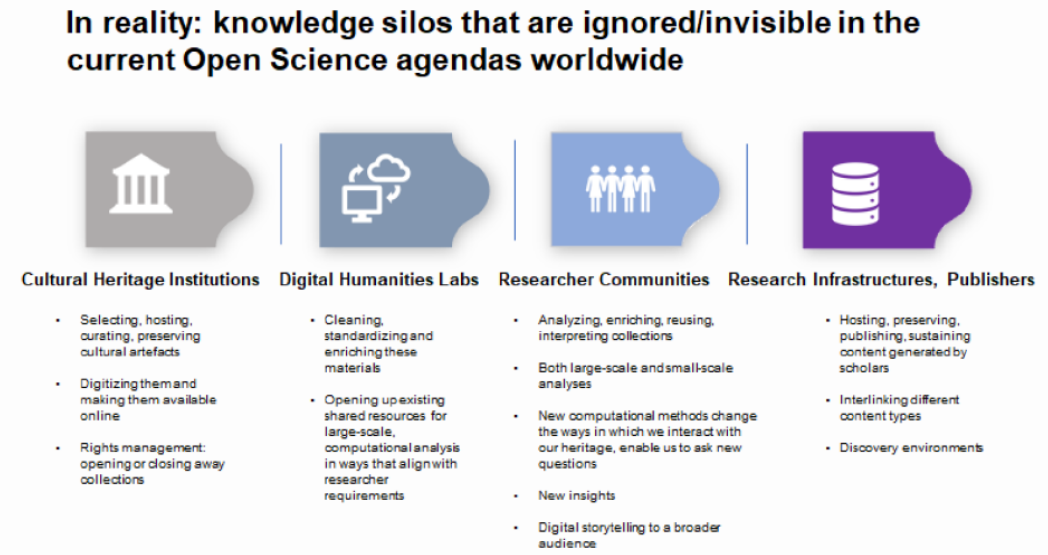
is far from being simple. To go beyond the scope of flagship institutions and national museums such as the Rijskmuseum, there is the example of an institution like the Belgian Royal Institute of Cultural Heritage (KIK-IRPA). Through its online platform, BALaT, the institution allows access to a large part of its over one million photographs of cultural and artistic heritage, among which there are many art works. One challenge in attribution in this context is that when images are used, the authors will most probably mention where the artwork is located, but rarely mention more detailed provenance information and will seldom indicate that they found the image through BALaT. Discussions were held in 2022 about the possibility of including a CC-BY licence instead of a CC0 licence in cases where Belgian federal scientific institutions share the works of their collections (at the same time additional funding was requested to be able to achieve this). These discussions have not yet been finalised in an official agreement.

Investing in and sharing detailed provenance data when making use of them as open (research) data is a cornerstone of FAIR-by-design data workflows in the domain and is, therefore, essential and important good practice for researchers. By doing so, they support their primary data providers and the cultural heritage institutions. A general clarification and thinning of the rights of an entire collection rarely seems possible within the available resources of an institution. In such cases, at least project-specific clarifications can be made, which then make partial collections available. This is a start, but unfortunately does not solve the jurisdictional problem concerning the complicated situation in collection institutions.

3.5.1. Why cultural heritage institutions should be involved in data management plans from the design phase, and how this could mitigate challenges coming from cross-sector knowledge silos

The example above clearly highlights one of the biggest shared challenges facing the diverse data workflows in the Arts and Humanities domain, that is, the data curation processes which fall along a natural continuum between a range of different actors, as depicted here. In reality, these different layers of curation, enrichment and analysis are separated by legal, institutional, infrastructural and even funding silos (as in many countries, these institutions belong to different ministries, fall under different legislative frameworks) and only in the rarest cases can they stay connected to each other.





Figures 6 and 7. Erzsébet Tóth-Czifra, *When open data starts with a handshake: tackling complexities of ownership, openness and silos in arts and humanities data workflows*. DOI: 10.25592/uhhfdm.12982

If implemented wisely, data management plans provide the opportunity to address, sort out, and eventually mitigate the legal, ethical, and technical challenges which come with such cross-sectoral knowledge silos. Used as project management tools rather than merely administrative tasks for fulfilling funder requirements, the process of data management planning can serve as a map of problems and as a roadmap for planned solutions in which challenges and obstacles to the responsible open reuse of source materials can be identified and addressed on time. This is when the rest of the research or project workflow is still flexible enough to modify or make decisions accordingly, and to dedicate the extra capacities or resources needed for making data available as FAIR.

The involvement of the cultural heritage professionals/curators of the collections to be reused (together with all other actors who have a role in data curation) is key. To frame

such joint data management planning, DARIAH-EU has published an open data guide for humanists (Edmond and Tóth-Czifra 2018) which, among others, provides a check--list of the most important questions to ask the archivist when visiting the archive, be it digital or physical; questions which pave the way for (FAIR and) sustainable publication and data sharing at the end of the project.

If questions concerning accessibility to, and reuse of, cultural heritage data during the project phase are not properly addressed prior to the start of the project, in the worst case scenario, it may fail. Even cultural heritage institutions themselves are not immune to the fact that due to legal changes in recent years concerning the rights of photographers, they may not be allowed to use older images which they have in their own archives if the photographer is unknown. In some cases, for example, an object lost or destroyed during a war, this may mean that existing pictorial information concerning the object is practically lost to research if the author is unknown.

3.5.2. Case studies and good practices emerging from projects encompassing the Working Group member's professional networks

3.5.2.1. Shared data management between the archive and the data archive in the Participatory Knowledge Practices in Analogue and Digital Image Archives project

The ongoing project Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA, also see discussion in Chapter 3.4) draws on three photographic collections and their associated metadata from the Schweizerische Gesellschaft für Volkskunde (SGV, archiv.sgv-sstp.ch), which has for several years already been archived at the Swiss National Data and Service Center for the Humanities (DaSCH). Since the data to be created during the project is intended to finally become part of the SGV's original assets, a close collaboration between the two involved organisations – SGV and DaSCH – is indispensable, because it involves modifying the original data model and, accordingly, the necessary cleaning of the original data, which creates expenses. The data management plan, written in 2019–2020, contains a general statement declaring that both organisations will collaborate and that the new data which is collected and produced will be added to the pre-existing collections. At the time of writing, a little less than half the lifetime of the project has passed; and it turns out that there are different visions – only part of the new data may be relevant for the Schweizerische Gesellschaft für Volkskunde, and so only these will be included within the original dataset archived

at DaSCH (ark.dasch.swiss/ark:/72163/1/0812). Thus another solution will have to be found for the remaining part of the new data or, in fact, the full set. Recently, the question of whose responsibility it was to clean the existing data to make them compliant with the new data model had to be solved. A point of current discussion is what should be regarded as part of the user interface, and what should be saved and become part of the data – since the PIA project is a citizen science project where images can be annotated by anybody and where users can upload their own resources, different ideas and visions may quickly arise. For instance, while the Schweizerische Gesellschaft für Volkskunde's focus of interest may be to identify the exact location of an image to enrich metadata through crowdsourcing, a user may, instead, focus on the people in the image and create interactive storytelling material or even memes. After more than a year out from the start of the project (February 2021), and as the project is financed until January 2025, it may be more than about time to resolve this issue. The interdisciplinary project team has recognised this problem and is currently discussing it, and is aiming to create a sustainable solution which continues beyond the end of the project. Thus, guidance within the process of data acquisition becomes relevant. From the perspective of a research data specialist, there is a lot to learn from this example – if one is involved in the pre-submission phase of such a project, we should try to settle in advance any questions about the differences in perspectives of the involved partners and their potential consequences by addressing them during this early stage so they can coexist. The following section is devoted to how this may be achieved.

3.5.2.2. The 'Data Reuse Charter', its associated tools for facilitating reuse agreements between involved parties, and their implementation

Such challenges have led several European organisations (such as APEF, CLARIN, Europeana, and E-RIHS) to come together and join forces under the governance of DARIAH to set up principles and mechanisms for improving the conditions for CH reuse. The vision we have in mind is to use the *Data Reuse Charter* to support cultural heritage institutions, infrastructure providers, and researchers to mutually clarify their goals at the beginning of a project, and to arrive at mutual reuse agreements concerning issues such as specifying access to data, provenance information, preferred citation standards, hosting responsibilities etc. which can go beyond the scope of what is specified in the licences.

As a first step in this joint effort, the team has established six basic principles which should be present and which should frame these exchanges. These are 1) Reciprocity,

in the sense that the different parties should share content and knowledge equally between each other, and mutually acknowledge each others' efforts based on partnership; 2) Interoperability, in the sense that cultural heritage data should be made accessible in a form which is suitable for research, and also digital research; 3) Citability, in the sense that both source materials and their enrichments should be easily and persistently citable, 4) Openness, which is obvious – quoting the famous axiom 'as open as possible, as closed as necessary'; 5) Stewardship, in the sense that the long-time preservation, persistence, accessibility, and legibility of cultural heritage data should be a priority; and finally 6) Trustworthiness, in the sense that as detailed as possible provenance information should be available on both sides, which explains how the data and how the resources have been processed. These principles are fully compliant with the famous FAIR principles and also probably with their younger sister, the CARE principles. As a second step, the principles went through multiple rounds of validation and even endorsement from both the researcher and cultural heritage sides through consultations, surveys, and workshops. The fact that, although the principles form a frame of reference which fits well with cultural heritage data exchanges, it is extremely difficult to translate into the everyday practices of institutions and research teams, something which has been repeatedly voiced in such consultations.

To bring the principles closer to everyday research and data exchange realities, in 2020, a subgroup of the original team involved in the charter designed a reuse agreement template for use between cultural heritage institutions and researchers which could serve as a starting point for FAIR-by-construction data management – right from the project planning/application phase.

In practice, the reuse agreement template can be flexibly applied in platform--independent ways. Institutions who sign the charter are able to use it (and expect to use such templates) in their own exchange protocols and publish it on their website, or can implement parts of it, for instance, a 'cite as' metadata field in their APIs. On the other hand, researchers can use the template to involve the cultural heritage sector in the initial stages of their data management planning process. As research data management planning is becoming an increasingly more common requirement among re-search funders, we need to raise the funders' awareness about the fact that such bi- or tri-lateral agreements and data reuse declarations between researchers, cultural herit-age institutions, and infrastructure providers are very important, and that DMPs in which cultural heritage data are involved should begin with this.

Below are a few implementation use cases of the Data Reuse Charter.

- Computational Literary Studies Infrastructure (inclusion of the charter in the project's DMP)

In the H2020 project, Computational Literary Studies Infrastructure, the reuse agreement template has been included as part of the data management plan to ensure that the reuse conditions of digitised literary works are sorted out in a sustainable, legal, and ethical manner, even if they are under copyright.

- LexSeal

The charter's six core principles also serve as the basis for a framework for the quality assessment of lexicographical datasets called Lexicographic Data Seal of Compliance. This work is affiliated with Elexis, the European infrastructure dedicated to Lexicography.

- DARIAH Campus

DARIAH Campus is DARIAH's training discovery platform. Under the *DARIAH Campus Reuse Charter*, we show how these commitments are reflected and implemented in the design and daily operations of the platform. For instance, under the principle of Reciprocity, the charter allows us to clarify our expectations regarding interactions between content creators, users, and curators as a mutual declaration of goodwill. Openness covers our licensing policy, while Trustworthiness stands for the provision of clear provenance information about the training resources we integrate into DARIAH-Campus and recognising the diverse range of contributor roles, also at the level of metadata, to clearly document who participated in the production process.

3.5.2.3. Data management in the Open Science in Arts, Design, and Music project

The OS-ADM project stands for Open Science for Arts, Design, and Music, and is led by the SUPSI (University of Applied Sciences and Arts of Southern Switzerland) with partners HES-SO (ECAL, Lausanne; HEAD - Genève; EDHEA, Valais), ZHdK (Zurich), HSLU Hochschule Luzern - Design & Kunst (Lucerne), BFH (Bern), and FHNW (Basel), with the support of swissuniversities and the involvement of DARIAH-EU Digital Research Infrastructure for the Arts and Humanities, the Swiss Artistic Research Network (SARN), the Swiss Design Network (SDN), and the Creative Commons. The project aims to support the implementation of the swissuniversities's open access action plan (www.swissuniversities.ch/en/topics/digitalisation/open-research-data/national-strategy-and-action-plan) in the interdisciplinary field of the arts (photography; visual and performing arts, such as dance, theatre and cinema), design (including sub-disciplines such as visual communication, industrial design, fashion design, and interaction design), and

music (including sound and aural arts). This field presents a series of complex issues which are related to the reuse and distribution of artwork and third-party content under copyright which is not accessible in the public domain and is subject to a series of restrictions. One can also note that these disciplinary fields produce a wide range of multimedia outputs. See OS-ADM (where the reuse agreement templates from the *Data Reuse Charter* will also be used) meta.wikimedia.org/wiki/Open_Science_for_Arts,_Design_and_Music

The structure of the project combines both centralised and decentralised approaches in order to meet individual needs, but ensures centralised management of the project, in which all partners are invited to regular meetings to inform, gather feedback, and involve all project members in the decision-making process. The coaching by the centralised team, which is oriented by the needs of the case study groups, is based on case studies proposed by the local teams.

However, there are challenges, including the open access publication of third-party content such as artworks; the heterogeneity of research outputs in the fields of art and design; difficulty in convincing scholars in the fields of arts, design, and media (ADM) of the importance of open science; and an unwillingness to open content for commercial use, as well as the fear of opening content which someone will then use for commercial purposes. To support researchers, this project integrates the reuse agreement templates from the *Data Reuse Charter* (halshs-03367459v2), which can help organise and settle cooperation between researchers and GLAMS.

In this project, the content and suggestions from other initiatives which support open science will be used for creating a space for brainstorming and discussion, which allows for reflection on the advantages of open science in the fields of art and design (i.e. open culture; collaboration with GLAMs; reuse of content from cultural institutions/archives, also for design, fashion, and new art works; increased visibility of research outputs; collaboration with open online collaborative projects such as Wikipedia, Wikidata, and Wikimedia Commons), for a focus on content produced within research which is financed by SNSF and European institutions (the explicit requirements of grantmakers), and for the application of institutional open access and open science policies.

One of the key issues is to find solutions for dealing with the legal restrictions placed on images related to cultural heritage management and their reproduction in print or digital publications. There is an urgent need to evaluate, and then educate and coach people in this field; this will be done through workshops and OA publications. The project outputs, including 13 open research case studies from the fields of art, design, and music as well as open science guidelines dedicated to this field, can be found at: meta.wikimedia.org/wiki/Open_Science_for_Arts,_Design_and_Music

3.5.2.4. Data management plan mandate and evaluation practices at BELSPO

Belgian Projects funded by the Belgian Science Policy Office (BELSPO) through Brain 2.0 have a strong DMP policy in place which concerns research institutions and cultural heritage institutions alike. The international evaluators of project proposals take into account the way in which the DMP is written, with the highest score for the most open data policy.[99]

3.6. Long-term archiving

Erik Buelinckx, Rita Gautschy, Francesco Gelati, Péter Király

3.6.1. Digital-preservation compliant file formats as the basis for long-term archiving

The common denominator in humanities projects such as digital editions, past exhibitions of cultural institutions, and archaeological excavations, is that the data should stay accessible, queryable, and interoperable. Selecting digital-preservation compliant file formats is a key step in guaranteeing the long-term preservation of research data. In other words, the long-term archiving of data in file formats which are not digital-preservation compliant, cannot be guaranteed. In particular, proprietary file formats are to be avoided; these types of files should be converted to open file formats, and the loss of information caused by this transformation should be regarded as the lesser evil. Another option is, of course, to only use recommended file formats for long-term archiving and to irreversibly delete the remaining files. These deletions should take place in the frame of archival appraisal procedures and should be (co-)performed and validated by archive staff.

There are several lists available of file formats which are recommended for archiving purposes. See the Catalogue of Archival Data Formats issued by the Swiss KOST-CECO Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen/ Centre de coordination pour l'archivage à long terme de documents électroniques (in German and French only) and the U.S. Library of Congress Recommended Formats Statement (in English only).

[99] www.belspo.be/belspo/openscience/opendata_DMPintro_en.stm

3.6.2. Data archiving and (digital) long-term archiving

Speaking provocatively, archivists regard long-term as being everlasting (which is not possible, but is something to strive for), while computer specialists are more interested in data archiving which should last approximately 10 years. Combining these two perspectives gives us a better insight into what is desired and what is doable. Everybody knows stories about the impossibility of reading old files due to obsolete formats or lack of support, bit rot, or similar issues. Realistically, it has to be kept in mind that, regarding long-term archiving, we are not all-knowing; so we need to be able to preserve things in a way which future generations are able to work with. It is this attempt to become as standardised and documented as possible in combination with good insight into the longevity of software and support which could become a solution.

ISO standard 14721, better known as OAIS (open archival information system), is seen as the ultimate model for long-term archiving (for an introduction, see Lavoie 2014). It provides both a conceptual model and a data model. The OAIS standard, together with the technology-independent PREMIS Data Dictionary for Preservation Metadata and the XML-based METS (Metadata Encoding & Transmission Standard) – both of which describe how digital preservation metadata should be structured – are the *gold standard* solution. METS, in turn, can be realised in several ways using different specifications: a good option is the E-ARK specification issued by the Digital Information LifeCycle Interoperability Standards Board (DILCIS Board). This system was explicitly recommended by the European Commission's Connecting Europe Facility eArchiving Building Block. Leading long-term preservation software (like the open-source Archivematica or the proprietary Rosetta) OAIS and METS compliant, which means there is a growing number of digital archives and digital libraries for long-term archiving in existence today (see, for instance, de Jong, Delaney, and Steinmeier 2013). Large research organisations should not only store their data in a data repository (data archiving), but also transfer their data to an OAIS-compliant facility (digital long-term archiving).

3.6.3. Long-term archiving and short-term financing

One example of the problems of long-term archiving can be found in the experience of the Belgian LTP-platform at BELSPO. Initially, the platform was created to preserve all the digital files created during the federal DIGIT-program (which is now in its fourth phase, 2019–2024). Given that the project will hopefully be continued, but with uncertainty about long-term financing, there needs to be a safety net. The LTP-platform itself does not require standardised files, although it is recommended, but has an obligation

to be able to return the files in exactly the same way as they were ingested should the funding stop; standardisation is up to the participating institutions. While this is rather simple for images, text, and other widely used formats, problems arise when files only exist in specific formats (often linked to a proprietary software). Some of these research data can be adapted to more standardised formats, but sometimes with a significant loss of information.

3.6.4. Long-term archiving of NoSQL/RDF/graph databases

Different approaches exist for the long-term preservation and accessibility of research data which are available in NoSQL form. As an example of this, the approach taken by the Swiss National Data and Service Center for the Humanities (DaSCH) will be introduced here. It focusses on a graph database solution which was developed for complex data obtained from the Humanities in Switzerland. The aim is to keep databases alive and, therefore have the data directly accessible and queryable. DaSCH developed a platform which consists of an RDF triplestore, an IIIF media server, and an API. To allow easier access to the data, a generic web application was developed. Each research project defines its own specific data model.

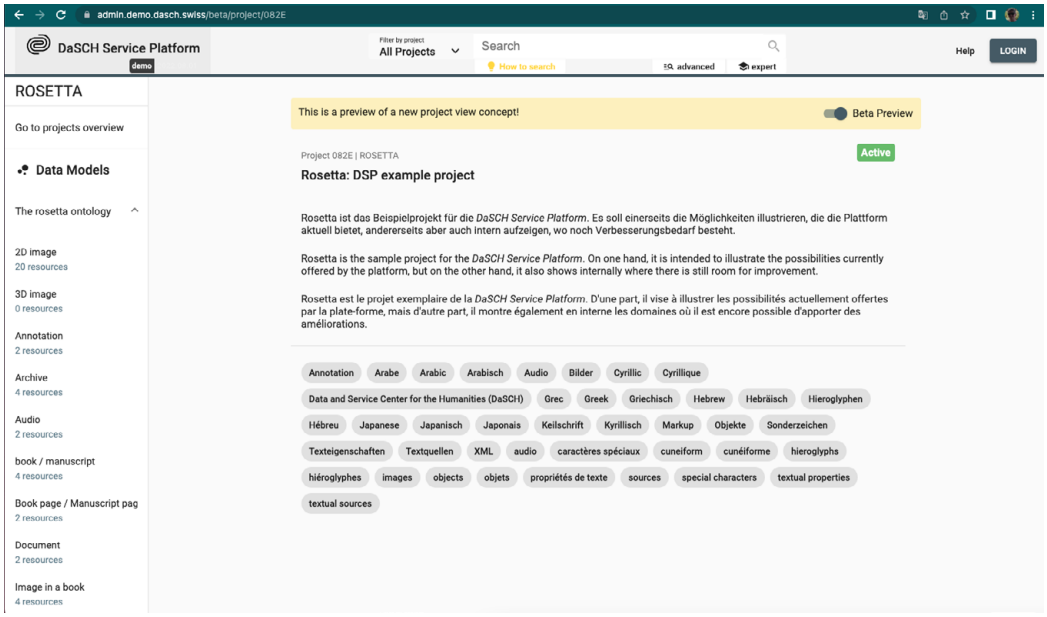
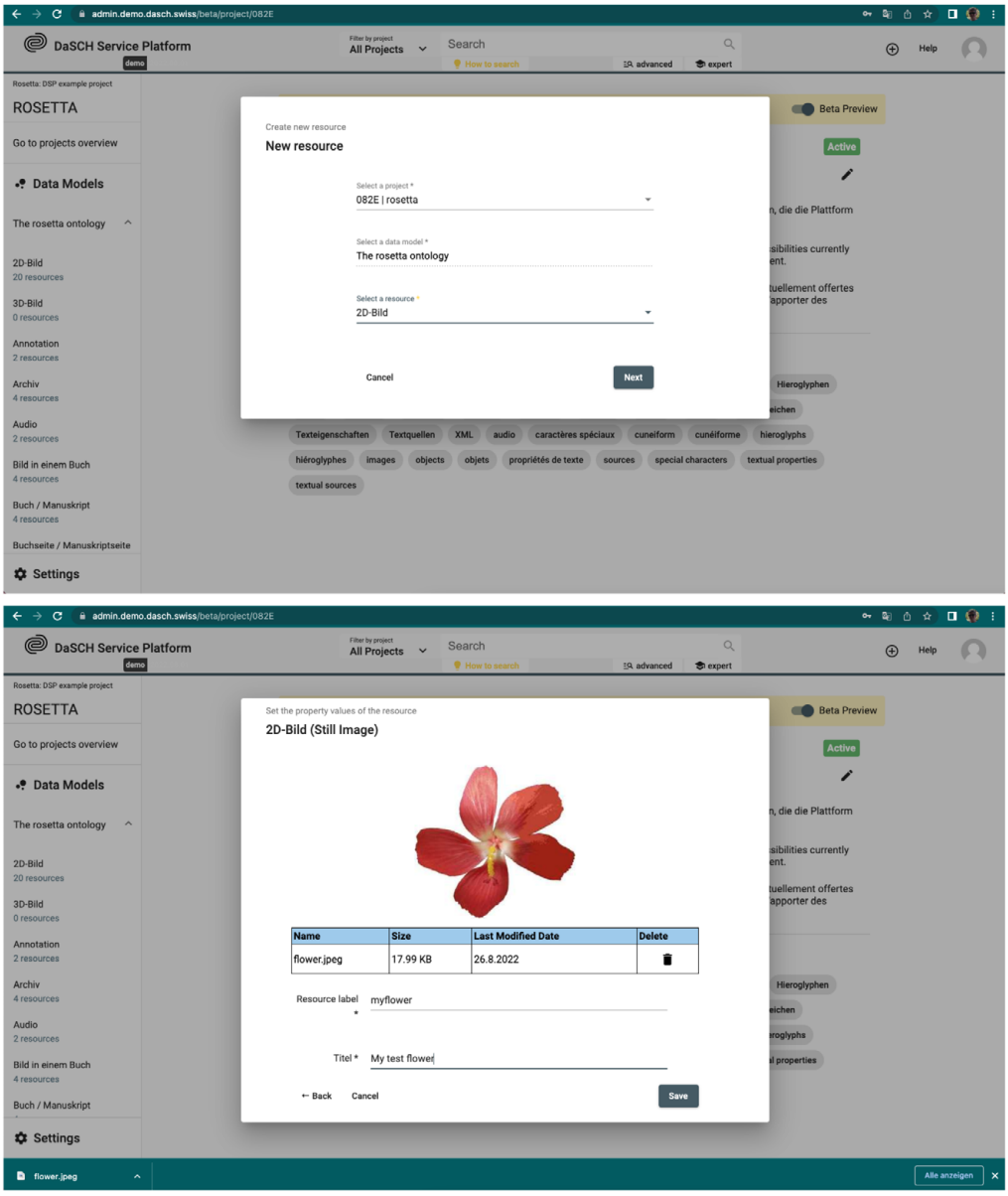


Figure 8. *The start page of the example project, Rosetta, on the DaSCH demo server. In the bar on the left you can see the resource classes which were defined for this project. Source: DaSCH Service Platform (DSP), admin.demo.dasch.swiss/project/082E.*

Depending on the research project, the data to be modelled for the graph database (cleaned and imported) are either already created on the platform via the generic web application, or they have to be exported from another tool or platform close to the end of the lifetime of the project.



Figures 9 and 10. *Illustrations of how to add data – in this case a 2D image. Source: DaSCH Service Platform (DSP), admin.demo.dasch.swiss/project/082E.*

In both cases the data may be modified later, and versioning allows for the retrieval of earlier states if necessary. Each resource within a project has its own permanent identifier, which allows for precise citations in publications.

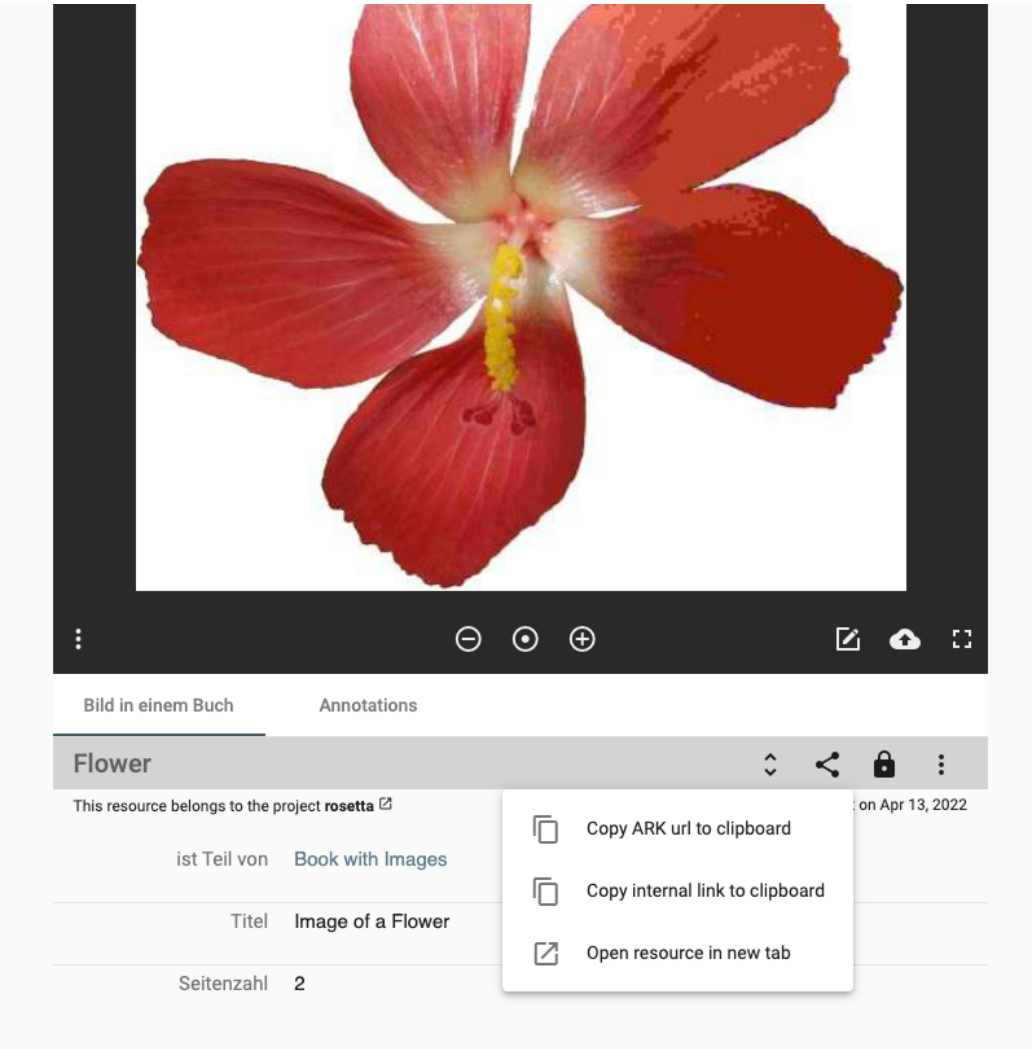


Figure 11. *Each object in the database has its own permanent identifier. Source: DaSCH Service Platform (DSP),* ark.dasch.swiss/ark:/72163/1/082E/4fj=PoIWQGqd5ISrbEbh0AH.

Various possibilities exist within the generic web application for searching the data. First, it is possible to limit the fulltext search to one specific project, or the search can be carried out across all projects. For more specific queries, the advanced search option is usually the better choice: by selecting a specific property or a combination of properties for the resource classes from the data model the search results can be narrowed down. The third option is the expert search – in this case it is necessary to write and submit a valid Gravsearch query in a pop-up window.

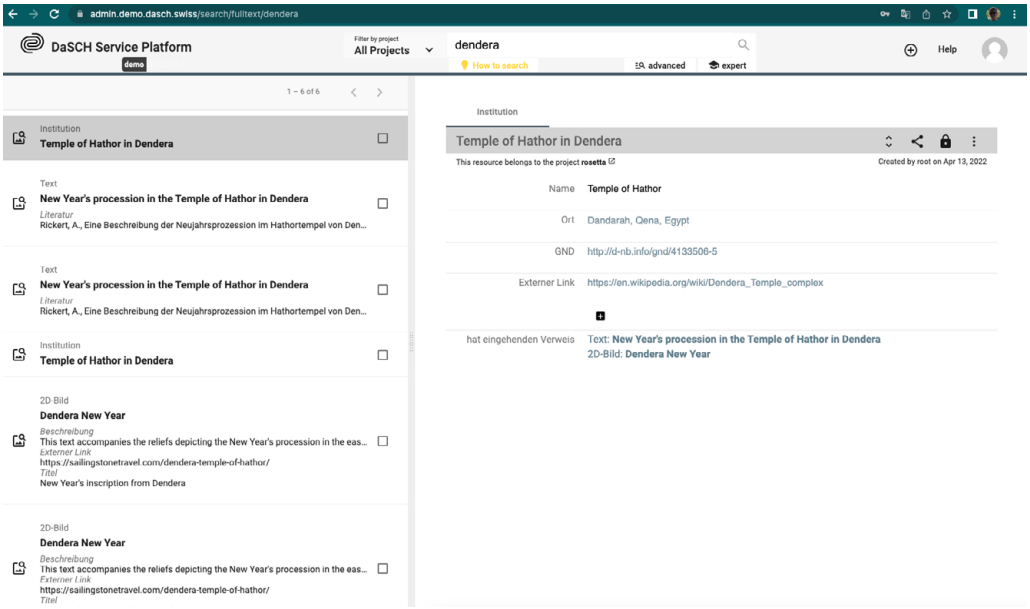


Figure 12. *Illustration of a full-text search across all projects on the server for the search term 'Dendera'. Source: DaSCH Service Platform (DSP), admin.demo.dasch.swiss/project/082E.*

3.6.5. Long-term archiving of relational (SQL) databases

In cases where the data have the form of a relational (SQL) database rather than NoSQL, the Swiss Federal Archives have developed the SIARD (Software Independent Archiving of Relational Databases) file format as a solution for its long-term archiving. A relational database, both in a proprietary file format like MS Access or Oracle, or in an open file format like MariaDB, can be transformed by means of freeware (e.g. the SIARD Suite, developed by the Swiss Federal Archives; or the DBPTK - Database Preservation Toolkit developed by the Portuguese company Keep Solutions) into a SIARD file without any loss of information (See: Naumann et al. 2022). Since the SIARD file format relies on other standards (XML, SQL:2008, UNICODE and ZIP64), it may be stored in a long-term archival facility and a copy of it may be reconverted into any of the supported file formats.

SIARD, like E-ARK, is maintained by the DILCIS Board. Yet its usage today remains quite limited; institutions which actively adopt it, like the Estonian National Archives or the Swiss Federal Archives, do not primarily work on humanities research data. As at June 2022, the Universitätsarchiv Hamburg had been planning to long-term archive the Universität Hamburg's Coronarchiv (a genuine example of humanities research data) by converting it into a SIARD file (see Gelati and Rau 2022). One year later (May 2023), a Docker container (a lightweight, standalone, executable package of software) seems to be a better solution, as the Coronarchiv is possibly too big to be transformed into SIARD.

3.6.6. Emulation

Emulation is a special method of preservation. Instead of migrating the file contents to future proof formats, emulation makes the reading devices backward compatible, so that modern machines can read old files.

Use case for the forensic approach:

Jürgen Enge - Heinz Werner Kramski - Susanne Holl: Friedrich Kittler's Digital Legacy. Digital Humanities Quarterly, 2017. Volume 11 Number 2.

I. Challenges, Insights and Problem-Solving Approaches in the Editing of Complex Digital Data Collections (digitalhumanities.org:8081/dhq/vol/11/2/000307/000307.html)
II. Friedrich Kittler and the Digital Humanities: Forerunner, Godfather, Object of Research. An Indexer Model Research (digitalhumanities.org:8081/dhq/vol/11/2/000308/000308.html)

Other relevant forensic articles:

John Durno: Digital Archaeology and/or Forensics: Working with Floppy Disks from the 1980s. The Code4Lib Journal, 34. 2016-10-25, journal.code4lib.org/articles/11986

Gregory Wiedeman: Practical Digital Forensics at Accession for Born-Digital Institutional Records. The Code4Lib Journal, 31. 2016-01-28, journal.code4lib.org/articles/11239

Julia Kim, Rebecca Fraimow, Erica Titkemeyer: Never Best Practices: Born-Digital Audiovisual Preservation. The Code4Lib Journal, 43. 2019-02-14, journal.code4lib.org/articles/14244

About the emulation process:
Dianne Dietrich, Julia Kim, Morgan McKeehan, and Alison Rhonemus: How to Party Like it's 1999: Emulation for Everyone. The Code4Lib Journal, 32. 2016-04-25, journal.code4lib.org/articles/11386

BIBLIOGRAPHY

Egil Bergenlind. 2017. 'Who Owns Your Personal Data under GDPR?' *DPOrganizer* (blog). 3 July 2017. www.dporganizer.com/blog/gdpr/who-owns-personal-data/.

Deniz Beyan, Oya, Chue Hong, Neil, Cozzini, Stefano, Hoffman-Sommer, Marta, Hooft, Rob, Lembinen, Liisi, Marttila, Juuso, & Teperek, Marta. 2020. 'Seven Recommendations for Implementation of FAIR Practice'. Zenodo. zenodo.org/record/3904140#.ZCk6mvZBy01.

Busa, R. 1980. 'The Annals of Humanities Computing: The Index Thomisticus'. *Computers and the Humanities 14* (2): 83–90. doi.org/10.1007/BF02403798.

Castro, Antonio Rojas. 2020. 'FAIR Enough? Building DH Resources in an Unequal World', August. hcommons.org/deposits/item/hc:32187.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2022. 'Archives and AI: An Overview of Current Debates and Future Perspectives'. *Journal on Computing and Cultural Heritage* 15 (1): 1–15. doi.org/10/gnsc2m.

Doran, Michelle. 2022. 'Seeing Shapes in the Cloud: Perspectives from the Humanities on Interdisciplinary Data Integration'. *Data Science Journal* CODATA 2022. www.tara.tcd.ie/handle/2262/98529.

Edmond, Jennifer. 2015. 'Tradition and Innovation in the Cendari Research Infrastructure'. *Review of the National Center for Digitization*, no. 26: 2–9.

Edmond, Jennifer et al. 2022. The Trouble With Big Data. Bloomsbury Publishing. www.bloomsbury.com/au/trouble-with-big-data-9781350239623/

European Commission. Directorate General for Communications Networks, Content and Technology. 2022. *Opportunities and Challenges of Artificial Intelligence Technologies for the Cultural and Creative Sectors*. LU: Publications Office. data.europa.eu/doi/10.2759/144212.

Gelati, Francesco. 2019a. 'Archival Metadata Import Strategies in EHRI'. In *Trust and Understanding: The Value of Metadata in a Digitally Joined-up World*, edited by R. Depoortere, T. Gheldof, D. Styven, and J. Van Der Eycken,15–22. Archives et Bibliothèques de Belgique - Archief- En Bibliotheekwezen in België 106. doi.org/10.5281/ZENODO.3746160.

———. 2019b. 'Implementing an Archival, Multi-Lingual and SemanticWeb-Compliant Taxonomy by Means of SKOS (Simple Knowledge Organization System)'. In *Proceedings of the Workshop on Language Technology for Digital Historical Archives - with a Special Focus on Central-, (South-)Eastern Europe, Middle East and North Africa*, 24–27. Incoma Ltd., Shoumen, Bulgaria. doi.org/10/gqjjvj.

Gelati, Francesco and Sönke, Rau. 2022. Presentation 'How to Archive a Corona Archive Project? The University of Hamburg's Coronarchiv and the SIARD File--Format'. doi.org/10.25592/UHHFDM.10494.

Gualandi, Bianca, Luca Pareschi, and Silvio Peroni. 2022. 'What Do We Mean by "Data"? A Proposed Classification of Data Types in the Arts and Humanities'. *Journal of Documentation*, December. doi.org/10.1108/JD-07-2022-0146.

Harrower, Natalie, Maciej Maryl, and Tímea Biro. 2020. 'Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities'. Digital Repository of Ireland. 2020. repository.dri.ie/catalog/tq582c863.

Kristina Hettne, Peter Verhaar (Centre for Digital Scholarship at Leiden University), Ben Companjen, Laurents Sesink, Fieke Schoots (Centre for Digital Scholarship at Leiden University, reviewer), Erik Schultes (GO FAIR, reviewer), Rajaram Kaliyaperumal (Leiden Universitair Medisch Centrum, reviewer), Erzsebet Toth-Czifra (DARIAH, reviewer), Ricardo de Miranda Azevedo (Maastricht University, reviewer), Sanne Muurling (Leiden University Library, reviewer). 2018. 'Top 10 FAIR Data & Software Things'. Zenodo. doi.org/10.5281/zenodo.2555498.

Jong, Annemieke de, Beth Delaney, and Daniel Steinmeier. 2013. *OAIS Compliant Preservation Workflows in an AV Archive: A Requirements Project*. Netherlands Institute for Sound and Vision, 5 September. publications.beeldengeluid.nl/pub/78.

Pawel Kamocki, Erik Ketzan, Julia Wildgans. 2018. *Language Resources and Research under the General Data Protection Regulation*. www.clarin.eu/sites/default/files/CLIC_White_Paper_3.pdf.

Király, Péter, Juliane Stiller, Valentine Charles, Werner Bailer, and Nuno Freire. 2019. 'Evaluating Data Quality in Europeana: Metrics for Multilinguality'. In *Metadata and Semantic Research, MTSR 2018, Communications in Computer and Information Science*. Edited by Emmanouel Garoufallou, Fabio Sartori, Rania Siatri, and Marios Zervas, 846:199–211. Cham: Springer International Publishing. doi.org/10/gqjjvh.

Kulczycki, Emanuel, Raf Guns, Janne Pölönen, Tim C. E. Engels, Ewa A. Rozkosz, Alesia A. Zuccala, Kasper Bruun, et al. 2020. 'Multilingual Publishing in the Social Sciences and Humanities: A Seven-Country European Study'. *Journal of the Association for Information Science and Technology* 71 (11): 1371–85. doi.org/10.1002/asi.24336.

Lavoie, Brian. 2014. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide* (2nd edition). Digital Preservation Coalition. doi.org/10.7207/twr14-02.

Malits, Andrea. 2020. Infrastrukturentwicklung für digitale Editionen am Beispiel der Universität Zürich: Herausforderungen, Erfahrungen und Perspektiven. *Bibliothek Forschung und Praxis* 44 (2): 202–209. doi.org/10.1515/bfp-2020-0023.

Maryl, Maciej, Marta Błaszczyńska, Agnieszka Szulińska, Anna Buchner, Piotr Wciślik, Iva Melinščak Zlodi, Jadranka Stojanovski, et al. 2021. 'OPERAS-P Deliverable D6.5: Report on the Future of Scholarly Writing in SSH'. Zenodo. doi.org/10.5281/zenodo.4922512.

Maryl, Maciej, Marta Błaszczyńska, Bartłomiej Szleszyński, and Tomasz Umerle. 2021. 'Dane badawcze w literaturoznawstwie'. *Teksty Drugie. Teoria literatury, krytyka, interpretacja*, no. 2 (March): 13–44.

Mons, Barend; Schultes, Erik; Liu, Fenghong; and Jacobsen, Annika. 2020. 'The FAIR Principles: First Generation Implementation Choices and Challenges'. *Data Intelligence* 2 (1–2): 1–9. doi.org/10.1162/dint_e_00023.

Mons, Barend; Neylon, Cameron; Velterop, Jane; Dumontier, Michelf; da Silva Santos, Luiz Olavo Boninob; Wilkinson, Mark D.h. 2017. 'Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud'. doi.org/10.3233/ISU-170824.

Naumann, Kai, Francesco Gelati, Kevin McMahon, and Oliver Watteler, eds. 2022. *Databases for 2080*: *Workshop Proceedings*. Stuttgart: Landesarchiv Baden-Württemberg. d-nb.info/1265590877/34.

O'Donnell, Daniel. 2022. 'Thinking about CARE Principles in the Digital Humanities. Why CARE May Not Be Only a Matter for Researchers Working with Indigenous Peoples'. July 25. doi.org/10.5281/zenodo.6903688.

Raemy, Julien Antoine. 2021. Applying Effective Data Modelling Approaches for the Creation of a Participatory Archive Platform. In Human Factors in Digital Humanities, ed. by Yumeng Hou. EPFL, Institut des humanités digitales, PhD Seminar. infoscience.epfl.ch/record/291219.

Schöpfel, Joachim, and Otmane Azeroual. 2021. 'Rewarding Research Data Management'. *WWW '21: Companion Proceedings of the Web Conference* 2021, April. doi.org/10.1145/3442442.3451367.

Daniel Skatz. 2017. 'FAIR Is Not Fair Enough'. Daniel S. Katz's Blog (blog). 22 June 2017. danielskatzblog.wordpress.com/2017/06/22/fair-is-not-fair-enough/.

Tasovac, Toma, Sally Chambers, and Erzsébet Tóth-Czifra. 2020. *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*. hal.archives-ouvertes.fr/hal-02961317.

Tasovac, Toma, Laurent Romary, Erzsébet Tóth-Czifra, and Irena Marinski. 2021. 'Lexicographic Data Seal of Compliance'. (Research Report) ELEXIS; DARIAH. hal.science/hal-03344267.

Tóth-Czifra, Erzsébet. 2020. '10. The Risk of Losing the Thick Description: Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem'. In *Digital Technology and the Practices of Humanities Research*, 126.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship' (Comments and Opinion). *Sci Data* 3, 160018 (2016). doi.org/10.1038/sdata.2016.18.