INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Metadata, Semantics and Ontologies

ISSN online: 1744-263X - ISSN print: 1744-2621
https://www.inderscience.com/ijmso

Proposal for a framework of contextual metadata in selected research infrastructures of the life sciences and the social sciences & humanities

Christian Ohmann, Maria Panagiotopoulou, Steve Canham, Petr Holub, Kurt Majcen, Gary Saunders, Maddalena Fratelli, Jing Tang, Philip Gribbon, Reagon Karki, Mari Kleemola, Katja Moilanen, Daan Broeder, Walter Daelemans, Pieter Fivez

Article History:
Received:
Accepted:
Published online: 30 August 2024

Proposal for a framework of contextual metadata in selected research infrastructures of the life sciences and the social sciences & humanities

Christian Ohmann*, Maria Panagiotopoulou and Steve Canham

European Clinical Research Infrastructure Network (ECRIN),
Bd Saint-Jacques, Paris, France
Email: christianohmann@outlook.de
Email: maria.panagiotopoulou@ecrin.org
Email: stevecanham@outlook.com
*Corresponding author

Petr Holub and Kurt Majcen

Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium (BBMRI-ERIC),
Neue Stiftingtalstrasse 2/B/6, Graz, Austria
Email: petr.holub@bbmri-eric.eu
Email: kurt.majcen@bbmri-eric.eu

Gary Saunders

European Infrastructure for Translational Medicine (EATRIS-ERIC),
De Boelelaan 1118, Amsterdam, The Netherlands
Email: garysaunders@eatris.eu

Maddalena Fratelli

Istituto di Ricerche Farmacologiche Mario Negri IRCCS,
Via Mario Negri, Milano, Italy
Email: maddalena@marionegri.it

Jing Tang

Faculty of Medicine,
University of Helsinki,
Haaatmaninkatu, Helsinki, Finland
Email: jing.tang@helsinki.fi

Philip Gribbon and Reagon Karki

Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP),
VolksparkLabs,
Schnackenburgallee 114, Hamburg, Germany
Email: philip.gribbon@itmp.fraunhofer.de
Email: Reagon.Karki@itmp.fraunhofer.de

Mari Kleemola and Katja Moilanen

Finnish Social Science Data Archive,
Tampere University,
Tampere, Finland
Email: mari.kleemola@tuni.fi
Email: katja.moilanen@tuni.fi

Daan Broeder

Common Language Resources and
Technology Infrastructure Network (CLARIN ERIC),
Utrecht University, BS Utrecht, The Netherlands
Email: d.g.broeder@uu.nl

Walter Daelemans

Department of Linguistics,
University of Antwerp,
Lange Winkelstraat 40, Antwerp, Belgium
Email: walter.daelemans@uantwerpen.be

Pieter Fivez

Antwerp Text Mining Centre (TEXTUA),
University of Antwerp,
Lange Winkelstraat 40-42, Antwerp, Belgium
Email: pieter.fivez@uantwerpen.be

Abstract: Usually, the focus of metadata annotation is on the research output rather than the context in which products were generated. The objective of this project was to develop a framework for contextual metadata, involving six research infrastructures (RIs) from two different domains. As a first step semi-structured interviews were performed to assess the current handling of contextual metadata. Then, these results were put into perspective with the main entities of research processes in general, leading to a framework for contextual metadata. From the discussion with the RIs and in alignment with the referenced literature, basic entities related to contextual metadata are defined and organised in a framework. In summary, a considerable amount of contextual metadata information is already covered by the RIs, however, not always explicit but implicit within text descriptions. The RIs involved see contextual metadata as necessary to improve replicability and reliability of research and FAIRness of data.

Keywords: contextual metadata; metadata schema; ontology; information model; EOSC; European Open Science Cloud; life sciences; social sciences & humanities; research infrastructures; interviews; knowledge graph; framework.

Reference to this paper should be made as follows: Ohmann, C., Panagiotopoulou, M., Canham, S., Holub, P., Majcen, K., Saunders, G., Fratelli, M., Tang, J., Gribbon, P., Karki, R., Kleemola, M., Moilanen, K., Broeder, D., Daelemans, W. and Fivez, P. (2023) 'Proposal for a framework of contextual metadata in selected research infrastructures of the life sciences and the social sciences & humanities', *Int. J. Metadata Semantics and Ontologies*, Vol. 16, No. 4, pp.261–277.

Biographical notes: Christian Ohmann has a Graduation in Mathematics (PhD), an interim examination in Medicine and a Habilitation in the field of Theoretical Surgery. He was the Head of the Coordination Centre for Clinical Trials (KKS) at the University Duesseldorf, Germany (1999–2014) and is now retired. Since 2004 he has been involved in ECRIN, currently as Chair of the Network Committee and of the Independent Certification Board. He has major competence and experience in the field of clinical research and clinical research informatics and has participated in a series of EU-funded projects.

Maria Panagiotopoulou received PhD degree in Biotechnology and she is a Senior Project Manager at the European Clinical Research Infrastructure Network located in Paris, France. She is experienced in the fields of cell biology, nanotechnology, toxicology, genetics, genomics, biomedical imaging and clinical research. Currently, she is coordinating ECRIN's data projects portfolio focusing on developing tools, good practices and guidelines that serve the European Clinical Research Community.

Steve Canham has 20 years' experience working with clinical research systems and data, after previous roles in healthcare and teaching. He worked on several data related H2020 projects, and has a particular interest in managing metadata to promote FAIRness of data.

Petr Holub is an Associate Professor of Computer Science at Masaryk University and Chief IT Officer (CIO) in BBMRI-ERIC, European Research Infrastructure Consortium for Biobanking and BioMolecular Resources, facilitating sharing of biobanking resources and health data for medical research. Since setup of BBMRI-ERIC, he has led the design and implementation of the

portfolio of its IT services. He is Co-founder of RationAI research group, focusing on rational and conservative application of explainable artificial intelligence to biomedical challenges. His research in computer networks and multimedia processing and in bioinformatics and applied artificial intelligence has lead to more than 80 research papers in established computer science, bioinformatics and medical informatics journals (including Nature Communications and Nature Review Genetics) and ranked conferences, and co-inventor of 2 patents. He has >1800 citations in Google Scholar (H-index 22). He has received Best Open-Source Software Award by ACM Multimedia SIG. He is a Co-founder of Comprimato company.

Kurt Majcen graduated in Information Technology and joined JOANNEUM RESEARCH for Database and Software Development in 1995. He built up experience with European projects in the cultural heritage area and with the Coordination of the European healthcare project NDSNET and led the team for the European project CLINICIP (Closed Loop Insulin Infusion for Critically Ill Patients). He set up a working group on assisting technologies, coordinated the AAL project ALICE and ran several corresponding projects (e.g., study on potential of AAL solutions for homecare, Austrian piloting regions). He also performs information management activities, e.g., introduction of document management, knowledge exchange and CRM systems. He conducted studies (feasibility study for a Digital Mediathek, pilot study about converging technologies for internet, TV and mobile devices), established a technical infrastructure and maintained ISO-14155 compliant procedures for data management within clinical studies. He joined BBMRI-ERIC as Project Manager and IT-Scientist in 2021.

Gary Saunders is the EATRIS Director of Digital Transformation. He is responsible for leading the EATRIS data strategy over the next scientific programme (2023–2026). As part of this work, he leads the EATRIS core data team and works with key data focussed initiatives such as the Horizon Europe Mission Areas, European Open Science Cloud and the 1+ Million Genomes. He joined EATRIS from ELIXIR where he was the Human Data Coordinator responsible for the implementation of the ELIXIR-wide strategy to enable responsible sharing of human data consented for reuse in scientific research. Previous to ELIXIR, he was based at EMBL-EBI where he was the Data Manager for the European Variation Archive (EVA), and the Database of Genomics Variants Archive (DGVa). He has a PhD degree in Bioinformatics from the University of Glasgow, UK, and has a background in Comparative Genomics.

Maddalena Fratelli is a biologist by initial training and became interested in bioinformatics and computational biology with the advent of the genomic era. Currently, she is Head of the Pharmacogenomics Unit at the Mario Negri Institute, is active in EATRIS, currently as Co-Chair of the Small Molecules Platform and participates in several EU-funded projects. She uses genomic and transcriptomic systems for the study of drug action, drug sensitivity or resistance and drug repurposing, with particular reference to the field of oncology. She is also interested in meta-science subjects such as open science, FAIR principles, data availability and re-use, and reproducibility.

Jing Tang is the Leader of the Network Pharmacology for Precision Medicine group and an Associate Professor in Medical Bioinformatics at the Faculty of Medicine, University of Helsinki, Finland. He received PhD degree in Statistics from the University of Helsinki. He is an Awardee of the prestigious ERC Starting Grant 2016, focusing on computational approaches to predict, understand and test personalised drug combinations in cancer. He has actively developed open-source systems medicine tools, supported by the EOSC-LIFE project and the European Infrastructure for Translational Medicine (EATRIS).

Philip Gribbon is Head of Discovery Research at the Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP) in Hamburg, where he leads several activities in data-driven target validation and early drug discovery. He is involved in Fraunhofer's programmes on targeted protein degradation and informatics initiatives for FAIR data and the application of machine learning methods to drug discovery and development. Recently, he has been appointed as Director General of EU-OPENSCREEN.

Reagon Karki received his Doctoral degree (PhD in Bioinformatics) from University Bonn/Fraunhofer SCAI in 2021. During his PhD, he worked in the field of Neurodegenerative diseases (Project: Aetionomy) and specialised in Alzheimer-Diabetes comorbidity through identification of shared genomics, genetics and mechanisms. His expertise and strengths are applications of systems biology which include knowledge and data integration, disease modelling, hypothesis generation and in silico validation. Currently, he works at Fraunhofer ITMP, Hamburg as a Researcher/Data Manager and contributes to projects such as PROXIDRUGS, BY-COVID, EOSC Future and AIOLOS.

Mari Kleemola is Development Manager at the Finnish Social Science Data Archive (FSD), Tampere University. She has 25+ years of experience in digital data preservation and open science, and her expertise areas include metadata, standards and certification of repositories. She has a Master's degree in economics. She has participated in several EU-funded projects building the European Open Science Cloud on topics related to metadata, interoperability, FAIR and repository certification and she is actively involved in CESSDA ERIC. Currently, she is a Member of the CoreTrustSeal Board, Co-Chair of the EDDI Conference Program Committee and leader of the CESSDA Trust and Landscape Working Group.

Katja Moilanen has 15+ years of experience in digital research data preservation and dissemination by working in Finnish Social Science Data Archive. Her main areas of expertise include data harmonisation, information and process modelling and metadata. She has been working on several metadata-related projects funded by EU and CESSDA (Consortium of European Social Science Data Archives). She graduated from Computer Science (MSc) and Social Sciences (BSocSci).

Daan Broeder has a background in Electrical Engineering, and has a long career working on research infrastructure, working in different capacities and tasks in various European and national projects. He led the metadata related work at MPI-PL TLA unit, for which he was the CTO and CLARIN research infrastructure. More recently, he was Coordinator for the EOSC cocreation SEMAF report on semantic mapping infrastructure. Currently, he is involved in the FAIRCORE4EOSC and OSTrails EU projects for CLARIN ERIC.

Walter Daelemans is Professor of Computational Linguistics at the University of Antwerp and leads the NLP group of the CLiPS research centre as well as the University of Antwerp's text mining core facility (TEXTUA). His research is on machine learning for natural language processing and its applications in social media analysis, clinical and medical text mining and computational stylometry.

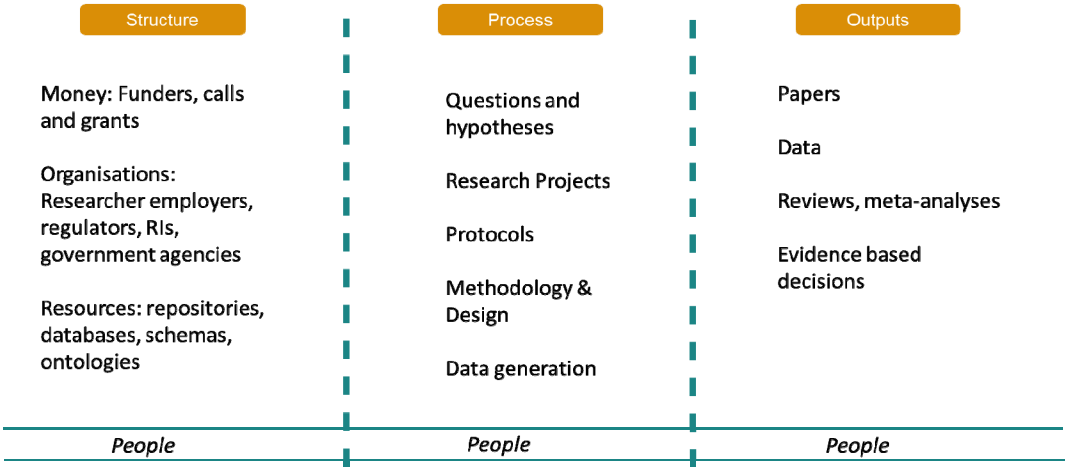
Pieter Fivez holds a PhD in Linguistics from the University of Antwerp, focusing on machine learning of semantic representations of biomedical text. Currently, he works as a Postdoctoral Researcher at the University of Antwerp, where he Coordinates the Antwerp Text Mining Centre (TEXTUA).

1 Introduction

The COVID-19 pandemic has generated a huge variety of research activities, studies and policies across both the Life Sciences (LS) and the Social Sciences and Humanities (SSH): examples include genomic sequencing, assays of immune response, clinical trials, population health analyses, exploring vaccine hesitancy, investigating the role of social media, public debate and economic analyses of the impact of public policy issues (e.g., lockdown measures, imposed face masking) (Pearson, 2021; Juul et al., 2020). Potential insights from combining the data and conclusions from these different forms of research are, however, made more difficult by the lack of a common metadata framework with which to describe them. The metadata landscape is heterogeneous and numerous domain-specific standards are applied, as recently demonstrated in the social sciences (Kleemola, 2020). The situation becomes even more complicated when data sharing is performed across broad disciplinary boundaries, as in this project, which spans life sciences, social sciences and humanities. Developing widely applicable metadata is a key part of rendering data more valuable, by allowing them to be more easily found and characterised, regardless of the discipline in which they were generated.

Usually, the focus of metadata annotation is on the research output (e.g., publication, report, data set). What is often missing is the characterisation of the context in which the research product was generated. Research design, approach, strategy, are heavily influenced by the researcher's epistemology and research philosophy (Al-Ababneh, 2020). The context, such as the type of the research (e.g., hypothesis testing versus hypothesis generating), the methodology chosen (e.g., experimental, survey, cohort, case study) and the research methods applied (e.g., type of sampling), are of major importance in understanding the data generated, and thus in supporting any secondary use of that data (Tobi and Kampen, 2018; Luff et al., 2015; Thiese, 2014). Contextual metadata are referring to a) data about the research process that generated the data, including descriptions of that process and the methodologies used and b) data about the 'inputs' into the research process – e.g., grants, people, organisations, regulators and research infrastructures and resources. In Figure 1 this is illustrated with a model originating from quality assessment.

Figure 1 Entities involved in research (structure, process, outputs) (inspired by Donabedian, 1980) (see online version for colours)



The problem is that different disciplines have vastly different ways of organising research activities, for instance because of differences in funding models and mechanisms, or in requirements for approval, and thus differences in how and when research is split into discrete activities and labelled. Therefore, major benefit is expected from better structuring and documenting contextual metadata. A higher replicability and reproducibility of research results can be achieved, and misconduct and research waste reduced, tackling one of the major problems raised in the past decade (Chalmers and Glasziou, 2009). What is needed is a common generic vocabulary, with which to describe, compare, assess and discuss metadata schemes and the contextual metadata they support.

1.1 Objective

The objective of this project was to develop a framework for a metadata model characterising contextual metadata, especially within research of the domains of six Research Infrastructures (RIs) from two thematic clusters. Involved were the following RIs from the LS: ECRIN (European Clinical Research Infrastructure Network), BBMRI (Biobanking and BioMolecular resources Research Infrastructure), EATRIS (European Infrastructure for Translational Medicine) and EU-OPENSCREEN (European Infrastructure of Open Screening Platforms for Chemical Biology) and from SSH: CESSDA (Consortium of European Social Science Data Archives) and CLARIN (Common Language Resources and Technology Infrastructure).

2 Methods

2.1 Semi-structured interviews of the RIs

As a first step in the process of assessing the handling of contextual metadata in the RIs, semi-structured interviews of nominated representatives of the participating RIs were performed as video conferences. During the interviews, objective aspects of the use of contextual metadata in the RIs domain were assessed as well as opinion-based and subjective views of the RIs about use and potential value of contextual metadata in their domain. The questions posed to the experts are summarised in Table 1 (see also appendix[1]).

Table 1 Questions posed to the experts (see questionnaire in the appendix)

Aspect	*Question*	
Objective aspects of the use of contextual metadata in the RIs domain	1.1	What does 'contextual metadata' mean to your RI? This question was divided into three parts: • How is the RI organising its services and tasks? What does that mean for contextual metadata that are directly applied within and by the RI? • What elements of contextual metadata of the resources/digital objects are modelled in the metadata schemas applied at your research RI (research organisations, researchers, services)? • What kind of contextual metadata are used in the domain represented by the RI?
	1.2	What services, protocols, standards, APIs are implemented in your RI to support harvesting of contextual metadata from outside (e.g., public or non-public API)?
	1.3	Are the contextual metadata in your RI already linked to a research process graph or is it planned to do so?

Table 1 Questions posed to the experts (see questionnaire in the appendix) (continued)

Aspect	*Question*	
Opinion-based and subjective views of the interviewees about use and potential value of contextual metadata in their domain	2.1	Do you believe that a greater generation and use of contextual metadata would be valuable enough to justify the additional effort that would likely be involved?
	2.2	From your viewpoint how could interoperability for contextual metadata between RIs be improved?
	2.3	What could be the best organisational framework for moving this work forward within EOSC (European Open Science Cloud)?

The study protocol and the interviewer guide were preregistered in ZENODO (Ohmann et al., 2022a, 2022b). The study protocol has been structured according to COREQ (Consolidated Criteria for Reporting Qualitative Research) (Tong et al., 2007).

During the interviews, the main research concepts applied within a RI were identified and a mapping of the entities identified to metadata schemas in the RI was performed. The interviews were recorded, and minutes were generated and approved by the interviewees. The final minutes from all interviews are available in the appendix.

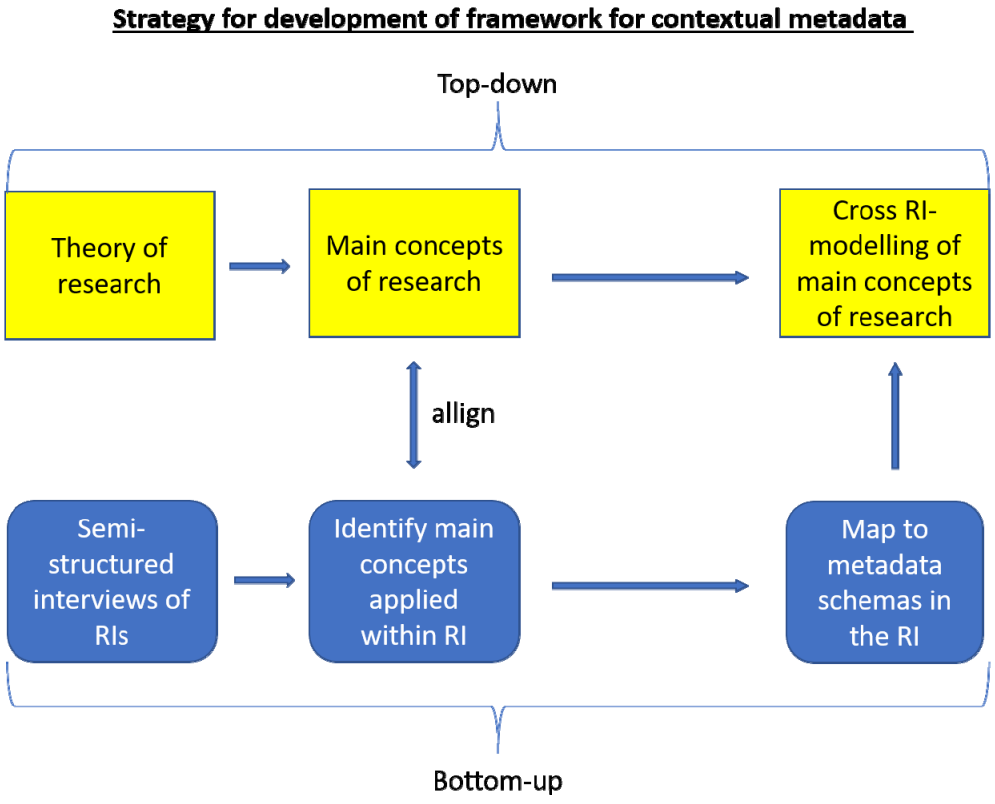
2.2 Synthesis and provision of a framework for contextual metadata

In the second part of this study, the results of the semi-structured interviews of the RIs were put into perspective by comparing the identified and mapped contextual metadata elements used in the RIs with the main entities of research processes as well as their relationships in general, derived from exploring existing information models/ontologies. The approach was used to synthesise the analysis as preparation for an overall conceptual metadata framework.

The overall methodological approach applied is shown in the Figure 2.

A first version of the full report from the study, including the interviews and the proposal for a framework for contextual metadata, was distributed to the project partners in March 2023. From the feedback received, a second version was produced and distributed in April 2023. Taking the feedback into consideration, a third and final version was released in June 2023. In addition, in April 2023 a virtual workshop with all participants was performed to discuss the report (version 2).

Figure 2 Methodological approach followed in the project (see online version for colours)



3 Results

3.1 Semi-structured interviews of the RIs

The semi-structured interviews were performed by interviewers from ECRIN between September 2022 and February 2023.

For better understanding of the complex and heterogeneous content, the main overall concepts for particular service(s) provided from the RIs were derived from the interviews and summarised in Table 2.

In one RI research projects are the target (ECRIN), three are dedicated to resource collections (BBMRI, CESSDA, CLARIN) and two RIs cover both aspects (EATRIS and EU-OPENSCREEN). For the majority of RIs the focus is to provide metadata with link to resources (ECRIN, BBMRI, CESSDA, CLARIN). EATRIS is focussing on standards and guidelines and EU-OPENSCREEN can be characterised as a central data hub. For four RIs contextual metadata are part of the research output (BBMRI, EU-OPENSCREEN, CESSDA, CLARIN), whereas for ECRIN and EATRIS the contextual metadata are of primary interest.

The detailed minutes of the interviews are provided in the full report of the project (see appendix). In the paper, only the main results are summarised (see Table 3).

Table 2 Main concepts in the individual RIs service(s) related to contextual metadata

RI service	*Basic entity*	*Primary goal to make findable*	*Basic entity characterised by*	*Purpose*	*Comment*
ECRIN (MDR – metadata repository)	Clinical study	Research project	Study characteristics and linked data objects	Metadata with link to resources	Contextual metadata primary focus
BBMRI Directory	Biobank	Resource collection	Sample, data collection	Metadata with link to resources	Contextual metadata part of research output
EATRIS (MICHA – Minimal information for Chemosensitivity Assay)	Chemosensitivity assay	Research activity/ Resource collection	Compounds, samples reagents, experimental design, data processing methods	Standards for annotation of drug screening protocols (guidelines)	Contextual metadata primary focus
EU-OPENSCREEN (EBCD – European Chemical Biology Database)	Assay	Research activity/ Resource collection	Compounds, targets	Central data hub	Contextual metadata part of research output
CESSDA (DC – Data catalogue)	Social science resources	Resource collection	Summary information, methodology, access	Metadata with link to resources	Contextual metadata part of research output
CLARIN	Language resources	Resource collection	Data collection protocol, data annotation guidelines	Metadata with link to resources	Contextual metadata part of research output

Table 3 Harvesting contextual metadata by the RIs (standards, services, APIs)

ECRIN	Used metadata schema	ECRIN metadata schema (Canham, 2023)
	Metadata services	ECRIN MDR
	List of elements of contextual metadata applied by the RIs	ECRIN uses the metadata provided during clinical study registration (e.g., on ClinicalTrials.gov). This includes information on data collection, time schedule, study content, people involved and participant population.
	APIs	Only internally available, public API planned
BBMRI	Used metadata schema	MIABIS (Minimum Information about Biobank Data Sharing) core 2.0 (MIABIS 3.0 under development) (Merino-Martinez et al., 2016)
	Metadata services	BBMRI-ERIC Directory PIDs (persistent identifiers) assigned to biobanks
	List of elements of contextual metadata applied by the RIs	BBMRI covers two major types of resources: Biobanks/collections linked to cohorts and clinical biobanks. The cohorts are usually research projects with textual description of the research question (research protocol). This is not the case for clinical biobanks where consented samples and data are added.
	APIs	MIABIS implemented with open source MOLGENIS software, Rest API to MOLGENIS available

Table 3 Harvesting contextual metadata by the RIs (standards, services, APIs) (continued)

EATRIS	Used metadata schema	Dependent on data type and selected repository
	Metadata services	MICHA (Different toolboxes currently under development)
	List of elements of contextual metadata applied by the RIs	For MICHA, the most important contextual elements for chemosensitivity assays have been standardised but each type of experiment is different and standardisation for other types of experiments remains a relevant need. For a typical drug sensitivity screening experiment there is a need to annotate five major components: compounds, samples, reagents, experimental design and data processing method.
	APIs	Available for MICHA
EU-OPEN-SCREEN	Used metadata schema	Uniprot (Universal Protein Database), ChEMBL (Chemical Database – EMBL), IC50 vals, Type, Organism, pChEMBL, CAS, SMILES (Simplified molecular-input line-entry system), Physicochemical props, Pathway IDs, GO (Gene Ontology) components, EFO (Experi-mental Factor Ontology) ids, synonyms
	Metadata services	The European Chemical Biological Database (EBCD). The COVID-19 Knowledge Graph (Karki, 2022). The Monkeypox Knowledge Graph (Karki et al., 2023)
	List of elements of contextual metadata applied by the RIs	The ECBD includes a standardised description of assays: an abstract in text format explaining how the experiment is formulated, information on the assay stage, assay type and an explanation based on the BioAssay Ontology is used to provide a standardised representation. ECBD contains information about the compounds, their structural format and calculated physical-chemical properties. Importantly quality control information is also provided.
	APIs	Available for ChEMBL, Uniprot for ECBD under development
CESSDA	Used metadata schema	CESSDA Metadata Profile (subset of CESSDA Metadata Model, which is subset of DDI – Data Documentation Initiative)
	Metadata services	CESSDA Data Catalogue
	List of elements of contextual metadata applied by the RIs	In the CESSDA Metadata Model (based on DDI), a lot of contextual metadata elements are applied. This covers, for example, main researcher, organisation, funder, contributors, topics, keywords, time-method, country, area, unit of analysis. The CMM has many more fields but not all metadata fields can be delivered in the CESSDA Data Catalogue (CDC). The European Language Social Science Thesaurus (ELSST) is recommended by CESSDA for data discovery across Europe.
	APIs	Available for CESSDA resources
CLARIN	Used metadata schema	CLARIN harvests metadata from its centres using OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) methods and from related projects and initiatives such as European
	Metadata services	Component Metadata Infrastructure (CMDI), Component Registry, Concept Registry, Virtual Language Observatory
	List of elements of contextual metadata applied by the RIs	Language materials are collected in various ways, covering the data elicitation method and the experiment type. There are plenty of metadata values available for describing a speaker or an assigner. Two elements of contextual metadata collected are: i) the data collection protocol and ii) the data annotation guidelines. In Natural Language Processing (NLP) technology, a lot of contextual metadata can be found in code and methodologies hosted on GitHub, usually in greater detail than is reported in scientific publications. In CLARIN, data collection protocols are available as resources pointed to in the CMDI metadata.
	APIs	Available for CDMI

All RIs interviewed in this study use their own and domain-specific metadata schemas and provide RI-specific services. The majority of RIs provide public APIs for their services, except for ECRIN and EU-OPENSCREEN, where such APIs are under development.

With respect to linkage of the RI-specific metadata to a research process graph, two RIs reported to be included as a source in EOSC Explore (EU-OPENSCREEN, CESSDA). For ECRIN mapping is ongoing and it may become relevant for EATRIS in the future. Two RIs stated that OpenAIRE needs to be updated and improved (ECRIN, CESSDA).

One specific question in the interviews was related to what the best organisational framework could be moving this work forward within EOSC. Integrating the results of the project into EOSC core services, onboarding to EOSC and registration in the EOSC catalogue is currently not seen as a topic for the majority of the RIs apart from CESSDA and EU-OPENSCREEN. Half of the RIs (ECRIN, BBMRI,

CESSDA) suggest exploring whether input into the EOSC Interoperability Framework could be useful. Input into the EOSC Association (e.g., Task Force Semantic Interoperability) is advocated by the majority of the RIs. In addition, a collaboration of EOSC with FAIRsharing and other resources was suggested.

From the opinion-based and subject views of the interviewees it can be concluded that contextual metadata are predominantly seen as necessary to improve replicability and reliability of research and FAIRness of data. The handling of contextual metadata is in different stages in the individual RIs but what is existing is generally not sufficient and needs to be improved. All six RIs except one (BBMRI) believe that a greater generation and use of contextual metadata would be valuable enough to justify the additional effort that would be likely involved. Half of the RIs (ECRIN, BBMRI, CESSDA) referred to the need to better understand the conceptual differences and similarities of contextual metadata between RIs. Crosswalks between metadata schemas may help but are not sufficient. A common research model across RIs is not propagated, a possible approach to improve may be to link services between RIs.

3.2 Synthesis and provision of a framework for contextual metadata

The analysis started with identifying the main entities of research and describing how these concepts are related and characterised by attributes. The idea was to have a generic approach to which the individual schemas of the RIs can be compared at and with each other. For the definition of the main entities and their relationships, existing information models and ontologies were taken into consideration. The work was based on four resources: BRIDG, ISA, the Core Ontology for Scientific Research Activities and the Model for Scholarly Research Activity. In BRIDG, a domain information model for translational and clinical protocol-driven research is provided (Becnel et al., 2017). To facilitate meaningful data exchange, BRIDG presents a common understanding of biomedical research concepts and their relationships with health care semantics. The open-source ISA framework and tools help to manage an increasingly diverse set of life science, environmental and biomedical experiments that are employing one or a combination of technologies (Johnson et al., 2021). The Core Ontology for Scientific Research Activities provides the design of a core ontology to deal with research activities (e.g., sampling and measurement: (Campos et al., 2019)). The Model for Scholarly Research Activity is a conceptual model for scholarly research activity, developed as part of the conceptual modelling work within the 'Preparing DARIAH' European e-Infrastructures project (Benardou et al., 2010). As the concepts used are fairly neutral with respect to different application domains, they can be reused to build ontologies for specific research domains. Table 4 summarises the approach.

Table 4 Basic entities and their relationships in resources from the literature

Resource	*Main entities*	*Relations between main entities*
BRIDG (Becnel et al., 2017)	'activity' 'study'/'experiment' 'research project'	'performer' performs 'activity' 'activity' performed in the context of 'study'/'experiment' 'study'/'experiment' is a 'research project'
ISA (Johnson et al., 2021)	'assay' 'study' 'investigation'	'assay' connected to 'study' 'study' is connected to 'investigation'
Core Ontology for Scientific Research Activities (Campos et al., 2019)	'research activity' 'sampling'	'sampling' is a 'research activity' 'measurement' is a research activity' 'research activity' is performed by 'research activity agent' 'research activity' has as principal 'research activity principal' 'research activity' adopts 'research activity procedure' 'research activity' uses 'device' 'research activity' locates 'geographic point' 'research activity' researches 'researchable entity'
A conceptual model for scholarly research activity (Benardou et al., 20210)	'research activity' 'research goal' 'actor' 'procedure' 'tools/service' 'method'	'research activity' follows 'procedure' 'research activity' has 'research goal' 'research activity' develops 'proposition' 'research activity' illustrates & represents 'concept' 'procedure' is assigned to 'actor' 'procedure' employs 'method' 'procedure' requires 'tool/service'

From the discussion in the project group and in alignment with the referenced literature, the following basic entities were derived and defined in this project.

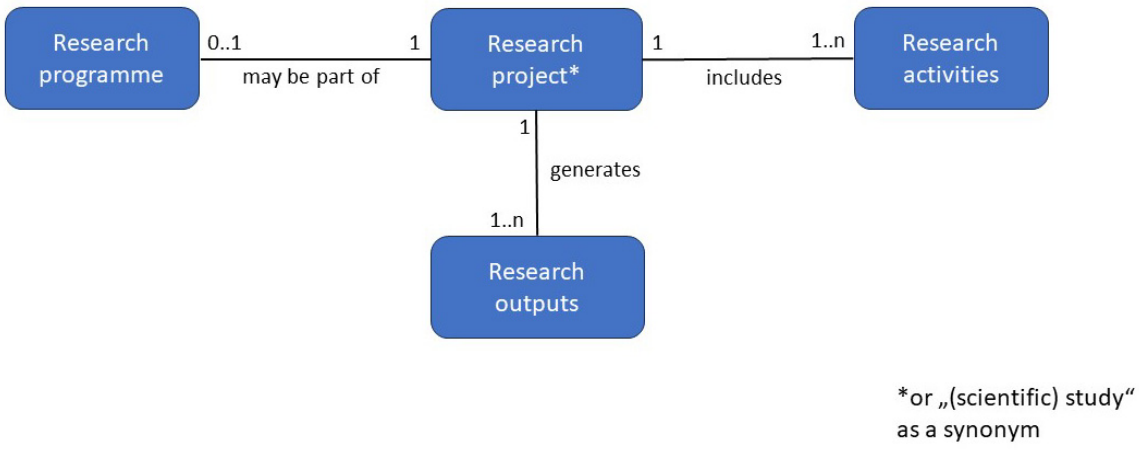
To avoid confusion, it is important to mention that the concept of 'research project' as it is defined above, is different from definitions by research organisations and funders as well as OpenAIRE (Open Access Infrastructure for Research in Europe). Structural aspects of projects are more tangible than the 'scientific goal', especially if there is no protocol associated with the project. There is therefore a tendency for many systems to focus on the structural aspects, as well as the equally concrete outputs. But that can leave a gaping hole in the middle – the research projects themselves, the processes by which the structural inputs led to the outputs. Any generic metadata scheme that claims to be about research needs to address this, hence the initial focus on defining a 'research project'. In the discussion this aspect is discussed in more detail.

To specify relationships between the main entities defined in Table 5 and for defining common attributes of the main entities, again input from the reviewed information models/ontologies has been used to come to a consensus in the project group. Figure 3 summaries the basic entities and their relationships in a class diagram.

Table 5 Definition of basic entities

Entity	*Definition*
Research project	A 'research project' is the set of research activities that attempts to answer one primary research question. The research project should be able to provide an answer to that question, where relevant with a statistical measure of confidence, or report that no answer was found. In the process of answering that question, a project provides a set of reportable and analysed data that has been derived from a single experiment or a set of closely related experiments and/or a set of observations on a defined set of related examples or subjects. A 'research project' can be a set of experiments in the lab, a clinical trial, a cohort, or an observational study in the life sciences or a study using interviews or questionnaires in the social sciences. Thus, a 'research project' may have 1 to n 'experiments'. However, the experiments must be closely related, in the sense that that they are all relevant to answering the single primary research question. 'study' (perhaps 'scientific study' would be better) and 'research project' are seen as synonyms. 'Study' on its own is too broad, but 'scientific study' can probably be assumed from the context. A 'research project' and / or a 'scientific study' can both be used to describe a unit of research work; one whose primary aim is to answer a particular scientific question. The work is 'scientific' in that it makes use of a systematic approach to generating or selecting, collecting and analysing data and does so with reference to existing knowledge and theories about the entities under study.
Research activity	A 'research activity' is a basic building block for a 'research project', in other words a 'research project' consists of 'research activities'. It has a defined goal and one or more actors that perform the activity. The goal is usually confined within a particular phase of the research project. Many activities have an object on which the activity is performed, and may have an output, something that is produced by the activity. Research activities may overlap in time. A research activity may include discrete procedures or tasks, that may be prescribed or described in detail, e.g., in the study protocol or within laboratory notebooks.
Research programme	'Research projects' may be bundled into 'research programmes'. A 'research programme' is a collection of research projects, that may be reported together but which were carried out at different times, often in a pre-planned sequence. A research programme may provide answers to one or more related research questions, each with, where relevant, a statistical measure of confidence or report that no answer was found. It may or may not be funded as a whole, or even by the same organisation. Examples are phase 1-4 studies about a treatment with the target of regulatory approval.
Research process	Research process: Most research projects can be divided up into 4 or 5 phases: a) Set-up, b) for interventional research only Carrying out the intervention, c) Data Collection, d) Data Analysis and e) Result and Resource Dissemination. These phases may overlap in time and may take on more or less importance in different types of studies. The research process usually follows a defined sequence with feedback loops.

Figure 3 Class relationships related to 'research project' (see online version for colours)



For the entities 'research activity' as well as 'research project', the following common attributes were defined:

In Table 7 the availability of common attributes for 'research activity'/'research project' in the literature resources is summarised.

As a next step, the availability and use of the basic entities and their relationships were investigated for the resources provided by the six participating RI service(s). Summary results are presented in Table 8, the detailed table is available in the appendix.

Table 6 Common attributes of 'research activity'/'research project'

Linked attributes to 'research activity'/'research project'
- has a goal
- has a location
- has actor
- has (researchable) subject
- has procedure
- has methods
- uses tools/services
- produces research output

Table 7 Common attributes of 'research activity'/'research project' in resources from the literature

Base class 'research activity' has the following attributes	*BRIDG High Level Concept Map (study) (Becnel et al., 2017)*	*ISA-Model (Johnson et al., 2021)*	*Core ontology for Scientific Research Activities (RA) (Campos et al., 2019)*	*Model for Scholarly Research Activity (RA) (Benardou et al., 2010) **
- has a goal	'Protocol' is linked to study	'Study' has 'protocol'	Missing (not at the core level)	RA has 'goal'
- has a location	'Study site' is linked to study	'Person' has 'role' has affiliation 'organisation'	RA locates 'geographical point'	missing
- has actor	'Person' is linked to 'subject' is linked to 'study'. 'investigator', 'study personnel', etc. are 'persons'	'Assay' has 'performer'	RA Is performed by 'research activity agent' Has as principal the 'research activity principal'	RA follows 'procedure' has 'actor'
- has (researchable) subject	'subject' is linked to 'study'	'Assay' has 'material'	RA researches 'researchable entity' (?)	missing
- has procedure	'Subject' has link to 'drug administration', 'procedure', 'observation'	'Assay' has 'technology type'	RA adopts 'research activity procedure' 'sampling', 'measurement' is RA	RA follows 'procedure'
- has methods	'Protocol', 'eligibility criteria', 'Arm' linked to study	'Study' has 'study design', 'Assay' has 'measurement type'	Not clear	RA follows 'procedure' employs 'method'
- uses tools/services	missing	'Assay' has 'technology platform'	RA uses 'devices'	RA follows 'procedure' requires 'tool/service'

Note: RA = 'research activity'*'research activities' can be nested and therefore do not represent one level of granularity.

Table 8 Availability of common attributes of a 'research project'/'research activity' in the RIs

'Research activity'/ 'research project'	*ECRIN MDR*	*BBMRI Directory (data collection)*	*EATRIS MICHA*	*EU-OPENSCREEN*	*CESSDA Data Catalogue*	*CLARIN CCR*
– has a goal	Implicit	Implicit	Implicit	Implicit	Implicit	Implicit
– has a location	Explicit	Explicit	Not available	Not available	Explicit	Explicit
– has actor	Implicit	Explicit	Not available	Not available	Explicit	Explicit
– has (researchable) subject	Partly explicit/implicit	Explicit	Explicit	Explicit	Explicit	Explicit
– has procedure	Partly explicit /implicit	Explicit	Explicit	Explicit	Explicit	Explicit
– has methods	Explicit	Explicit	Explicit	Explicit	Explicit	Explicit
– uses tools/services	Implicit	Implicit	Explicit	Explicit	Not available	Explicit
– produces research output	Explicit	Not available	Not available	Not available	Explicit	Explicit

Note: *Not available*: Not included in metadata elements and not derivable from other metadata information
Explicitly available: Metadata element explicitly available in the metadata schema
Implicitly available: Metadata element not explicitly available and but may be created from free text or derived from other metadata elements

In the considered Ris there is the following relation to 'research project'/'research activity':

- *ECRIN MDR*: Clinical study as 'research project' described in 'study schema'
- *BBMRI*: Sample collection as a result of a 'project'
- *EATRIS*: 'Chemosensitivity assay' as 'research activity'
- *EU-OPENSCREEN*: 'Assay' as 'research activity'
- *CESSDA*: Collection of data sets as result of a 'project' (called 'study')
- *CLARIN*: Language resource as result of a 'project' (described in 'project description').

For three RIs, 'research project' or 'research activity' are the main entities (ECRIN, EATRIS, EU-OPENSCREEN). For these RIs information on most of the common attributes of a 'research project' or a 'research activity' are available, however, not always explicitly named but included somewhere in the metadata. In the other three RIs, primarily resources are described that have been derived from 'research projects' or other types of projects. In BBMRI the entity 'study' is already foreseen in the data model but not yet implemented and a model has already been constructed linking 'study' with the other entities 'biobank' and 'sample collection' (Scapicchio et al., 2022). For CESSDA and CLARIN, information about a 'research project' underlying the sample or resource collection, is often available but distributed over different parameters of the metadata documentation. Here, primarily data sets that are outputs of projects are described. In summary, a considerable amount of contextual metadata information is already covered by the RIs, available partly explicit, partly implicit.

4 Discussion

4.1 Definition of 'research project'

Defining a 'research project' is central to exploring and developing a generic metadata system for research. It is also difficult, because of the various ways in which research is structured, funded, reported (etc.) in different disciplines, which therefore each have their own nuanced interpretation of 'research project'. For example, for research organisations and funding bodies a 'research project' is an administrative unit that has a budget, typically made up from one or more grant awards, under the control of a principal investigator and a project manager and with specific aims and objectives. However, the relationship of projects with funding can be complex. Whilst there may be a 1-to-1 link between a grant and a 'research project', in many cases a grant or fund may contribute to several projects, and an individual project may use money from different sources, including central or institutional funding. Funding is clearly a structural aspect of any 'research project', along with staff, buildings, tools and equipment, but that structure is not the 'research project' itself (though it might be how a funder most easily conceptualises it). The structure enables the activity, it provides a research project 'framework' or 'kit', but it is not the 'research project'.

In our concept, 'research projects' may be bundled into 'research programmes'. For research organisations and funding bodies, this is often a synonym for 'funding programme', to which principal investigators can apply for project funding. Again, as discussed above, a funding programme is clearly a structural aspect for a series of 'research projects', often dedicated to specific research domains and/or research questions formulated in a call, but it does not constitute the 'research projects' as such.

The definition given in this paper characterises a 'research project' as a goal-oriented process, one that can be applied to all scientific disciplines. It does not try to capture or distil the various alternative definitions or notions of 'research project' that currently exist. It is instead an attempt to define a core concept that can be used as the main 'unit of work' within research. Inevitably the definition retains a degree of ambiguity – in deciding what a 'primary research question' is, though the expectation is that within any discipline custom and usage should be able to provide a reasonably clear idea. But it does offer a simple way of differentiating the 'research project' from both its component activities and any broader programme(s) of research of which it is part, and it does provide sufficient flexibility to meet a variety of use-cases.

4.2 Machine-actionable metadata

Increasingly, funders, data managers/stewards and a variety of consumers of data and metadata are encouraging researchers to generate metadata in ways that can be retrieved, read and processed largely or entirely by computers. There is an important distinction to be made between the terms 'machine-readable' and 'machine-actionable'. Machine-readable metadata means a machine can easily parse the metadata stream into key value pairs; anything in XML, CSV, JSON, etc., would satisfy that condition. Machine-actionable is a claim that the structure is deep enough to allow inferences. So, for example, metadata for a research output that encodes 'creator' as an arbitrary string literal property of the output would be machine-readable but not machine-actionable. To be actionable, the metadata would need to represent 'creator' (wholly or in part) as a relationship to an agent entity/record, typically by supplying a persistent, unique identifier for that agent, thereby allowing a machine to traverse the graph from the output to the agent to other outputs, and so on. Machine-actionable metadata is necessarily highly structured and highly granular – sufficient for each data point of interest to be clearly and separately identified rather than being buried within a textual description (which would require human input for its interpretation). The data elements, the relationships between them, and the allowable values each can hold all need to be explicitly defined. Whether this proposal of a framework for contextual metadata could be transformed into a machine-actionable metadata model, depends primarily on the machine-actionability of the metadata schemas of the individual RIs. The situation is complex and difficult, however, because so much depends on the source systems in use within a domain and the willingness of metadata creators to agree on, learn and then use a highly structured descriptive system. Resources and time in the project were too limited to come to a full machine actionable metadata model in the project lifetime. What was possible in the project was firstly to create greater awareness about contextual metadata in different RIs from different domains and secondly to demonstrate the need to invest more resources in this field if interoperability is to be increased. Thirdly, the use of, and gaps in, contextual metadata in the RIs was assessed in semi-structured interviews and workshops. Fourthly, the availability of specific contextual metadata information in the RIs linkable to specified and common entities and attributes in the framework was assessed. This project represents a good starting point but much more is needed to progress. Making metadata 'machine actionable' needs, primarily, less ambiguous and more structured metadata, with entities, categories and ontologies all unambiguously labelled and clearly defined. Here, additional and extensive input from the scientific community would be required (Batista et al., 2022).

4.3 Research graphs

There are substantial gaps between the concepts of a research graph and its components and the metadata systems currently in use in the RIs. Nevertheless, it may be that a graph-based data structure, which would echo the actors – processes – objects structure more closely, would be a better approach to storing data about the different aspects of 'research activity'. Here it is explicitly stated that research graphs can connect entities, one aspect of an ontology, but they have little to say about the underlying vocabularies (another important element of an ontology). So, agreement on the terminology is another important aspect to be considered. It should be explored which knowledge (research) graph-based data structures are available and whether to use them gives advantage over traditional approaches. Most widely used and of major relevance for the EOSC is the OpenAIRE research graph data model. The OpenAIRE Graph includes metadata and links between scientific products (e.g., literature, data sets, software and 'other research outputs'), organisations, funders, funding streams, projects, communities and (provenance) data sources – the details of the OpenAIRE Research Graph Data Model can be found in Zenodo.org (Manghi et al., 2019). As such, the OpenAIRE graph already includes some of the basic entities to model contextual metadata (e.g., funder, project, organisation). Unfortunately, the research process, covering 'research projects' and 'research activities' as described in this paper, is not modelled explicitly. Within the FAIRCORE4EOSC project, an EOSC Research Discovery Graph (RDGraph) will be developed that will become a flexible and federated EOSC search service across EOSC repositories that extends EOSC Research Catalogue making it compatible with the specifications provided by the RDA's (Research Data Alliance) 'Open Scientific Graph for FAIR Data' working group and incorporating additional entities like the Research Activity Identifiers (RAiDs). In its core, it is based on the OpenAIRE Research Graph, and it will become exposed through its APIs and data dumps. The Scholix Framework (SCHOlarly LInk eXchange) is another high-level interoperability framework for exchanging information about the links between scholarly literature and data, as well as between data sets (Burton et al., 2017). The aim of the Scholix initiative is to find consensus on solutions to facilitate the exchange of information about semantic links between data sets and literature objects, which is key for reusability and reproducibility of science. The focus of Scholix is on the link between 'research outputs' of a project rather than on the contextual metadata of the 'research process'. Research

Graph schema is an accessible meta-model for connecting research objects, such as researcher, publication, data set, grant and organisation. This schema is designed to provide a practical approach to construct large scale graphs from a distributed network of scholarly works. It is built upon the main entities, researcher, publication, grant, organisation and research data. Again, the contextual metadata elements of the 'research process' ('research activity', 'research project') are not explicitly covered. Finally, the Scholarly Knowledge Graphs for EOSC (ORKG) should also be mentioned (Stocker et al., 2022). In summary, there seems to be major potential in using knowledge (process) graphs to assess and document contextual metadata. Some of the current approaches are promising but essential entities that characterise the 'research process' are still not taken adequately into consideration.

4.4 Identifiers

Within any research graph or research metadata system the allocation of a Persistent Identifier (PID) to the 'research activity' itself – to each distinct 'research project' – would seem an obvious starting point. In reality, however, while there are PIDs for many outputs (DOIs, Pubmed identifiers, etc.) and some inputs (e.g., Grant IDs, ORCID personal identifiers, ROR organisational identifiers) the central research activity itself, in most domains, is rarely allocated a PID. Most research, however, appears to be described retrospectively if and when it results in a published paper, otherwise remaining invisible outside of its source organisation, even to those that funded it. The lack of PIDs for 'research projects' produces a significant central hole in any research graph and breaks the PID chains that could otherwise be constructed all the way from inputs to outputs. Within the FAIRCORE4EOSC project, an EOSC RAiD service will be integrated within the EOSC Marketplace. RAiD provides persistent, unique and resolvable identifiers for 'research projects' based on the global Handle System. RAiD also collects descriptive information about the project activities and records these in a 'metadata envelope'. The EOSC RAiD will mint PIDs for research projects, which will allow authorised EOSC users and services to manage information about project-related participants, services and outcomes. The EOSC RAiD will thereby collect the relationships between research objects, which enriches analysis, tracking and reporting (including EOSC service utilisation), and indirectly supports reproducibility and extends the ability to discover research entities in the EOSC RDGraph/PIDgraph. Since the EOSC RAID is currently under development, it certainly would make sense to check whether the main entities and relationships identified in this project will be adequately covered in the planned services. For that reason, this paper is intended to be used as input into the FAIRCORE4EOSC project.

4.5 Provenance

Provenance is another aspect to be discussed in relation to contextual metadata. For the purpose of the paper the consensus definition of provenance from W3CPROV-DM standard is used: Provenance is information about entities, activities and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. (https://www.w3.org/TR/2013/REC-prov-dm-20130430/). The concept of prospective provenance is not that common and if used, it is usually based on assumptions and expectations, which are however not necessarily met practically (Lim et al., 2010). In practical experience, the prospective provenance can be based on provenance templates (Vasa et al., 2017; Moreau et al., 2018) describing standard operating procedures, e.g., as a part of laboratory notebooks (Schröder et al., 2022). The actual retrospective provenance can be captured as a link to the standard operating procedure and documentation of deviations from it.

Provenance provides a critical foundation for assessing authenticity, enabling trust and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance. In consequence, there is much overlap between provenance and metadata. In ISO 23494 Biotechnology, a provenance information model is defined for biological specimens and data (Wittner et al., 2021). The purpose is the standardisation of provenance information for the biotechnology domain, and it covers the whole process chain, from the source of biological material, through its processing, analysis and all steps of data generation and processing. The standard covers the provenance of sample acquisition, processing, transport and storage, of data generation and data storage and processing, activities which can be clearly seen as 'research activities' in our notation. Provenance does not, however, include all aspects of contextual metadata. It is primarily a retrospective approach, being applicable only when an object, whether digital or real, has been created. What is missing is the link to the generating 'research project' (when available) or context, and the main entities linked to the genesis and planning of the project (e.g., goal, hypothesis study type). Provenance and other contextual metadata are complementary, and a close link would be of major benefit.

4.6 Standards

There are several standards available, potentially relevant for better documentation of contextual metadata. The Data Catalogue Vocabulary (DCAT) will become of major importance. DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the web (https://www.w3.org/TR/vocab-dcat-3/). The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on DCAT for describing public sector data sets in Europe. Its basic use case is to enable cross-data portal search for data sets and make public sector data better searchable across borders and sectors (https://op.europa.eu/en/web/eu-vocabularies/dcat-ap). For the planned European Health Data Space (EHDS), the standard for a common descriptive metadata model will be based on a health DCAT-AP extension (https://ehds2pilot.eu/). It should be explored whether and

how contextual metadata as described in this paper could be reflected in the envisaged extensions of the DCAT standard.

With respect to publications, reporting guidelines have been formulated for different types of studies ('research projects'), that are required more and more by publishers. These guidelines are accessible through the EQUATOR (Enhancing the QUAlity and Transparency of health Research) network and cover randomised trials, observational studies, systematic reviews, study protocols, diagnostic/prognostic studies, case reports, clinical practice guidelines, qualitative research, animal pre-clinical studies, quality improvement studies and economic evaluations (https://www.equator-network.org/). The guidelines refer to contextual metadata of a research project, from which the publications were generated. If the guidelines are correctly applied in a publication, essential contextual metadata can be found but a link to the 'research project' as such is usually not available, and the standardisation of the information related to individual parameters of the guideline is not strictly. In addition, there may be inconsistencies if different papers related to one specific 'research project' are published in different journals. A specific type of a 'research activity' of a 'research project' is the creation (and maintenance) of a Data Management Plan (DMP). The DMP describes the data that is used and produced during the course of 'research activities'. The DMP Common Standards Working Group has developed an application profile that allows to express information from traditional DMPs in a machine-actionable way. It allows for automatic exchange, integration and validation of information provided in DMPs. The metadata application profile provided by RDA covers essential elements of contextual metadata, such as 'project', 'funding' and 'contributors'. The entity 'project' covers 'description', 'start', 'end', 'funding (ID-identifier, type), funding status, grant ID (identifier, type) and 'title'. The application profile has been implemented, e.g., in ARGUS, an open extensible service integrated within OpenAIRE and freely offered for use through the OpenAIRE Service catalogue and EOSC Catalogue. If implemented and applied, the common standard for machine actionable DMPs may significantly contribute to better link contextual metadata to projects and their outputs (e.g., data set), however, it does not solve the issue of defining a 'research project' as a separate explicit entity with links to different 'research activities' (e.g., DMP). So, it is certainly a step forward but alignment between these and other activities is needed to stepwise improve the documentation of contextual metadata.

The work performed in this project should be closely linked to the Metadata Schema & Crosswalk Registry (MSCR), currently under development in FAIRCORE4EOSC. The MSCR allows registered users and communities to create, register and version schemas and crosswalks with PIDs. The published content can be searched, browsed and downloaded without restrictions. The MSCR will facilitate the transformation of data from one schema to another via registered crosswalks. The framework for contextual metadata and crosswalks to other metadata schemas can be shared with the community for reuse and extension, thus improving step by step the possibility to explicitly characterise 'research projects' and 'research activities', to provide links between 'research projects' and the 'research output' and finally to improve interoperability cross-RIs and domains.

4.7 Other aspects

Access control (and the other sensitive data metadata points) should always be included within a generic metadata schema, even if they are only applicable to certain data set types – most obviously that concerned with human subjects or human derived material. Ideally, they would be *compulsory* for such data. If that was the case, human data that was made publicly available should then also have metadata that described how and why this was the case (e.g., because it had been de-identified, and / or was anonymised) to give a fuller picture of the data and more confidence in its safe re-use. Such data that was not publicly available should always have clear information about the circumstances, if any, that it might be made available, to whom and in what contexts, and give the details of any application process. It should be explored whether and how access control and other aspects not discussed in this paper need to be related to contextual metadata.

The biggest hurdle implementing any rich metadata collection is the cost-benefit calculation. In order to change behaviours, the benefits must not only clearly outweigh the costs but the researchers who participate need to experience a net decrease in workload as a result. Certainly, there is a need for automation to minimise the effort needed at an individual level to implement this. Use of tools to collect and structure the metadata is not only 'interesting' but vital if this proposal is to succeed. One option to be explored should be AI (Artificial Intelligence) — or ML (Machine Learning) algorithms to support automatic (or at least semi-automatic) detection of contextual metadata from text documents linked to resources. With such a text mining approach, if successful, missing entities of contextual metadata could be complemented. It is not easy to use AI/ML in this field due to the multilingualism and the potential misinterpretation of terms. Often there are different meanings between scientific disciplines and a common backbone for the application of AI/ML is difficult to achieve (David et al., 2022). It would certainly be of major interest to perform a detailed but painstaking analysis of the costs (time, money, corporate effort, individual effort) to implement collection of rich, standardised, contextual metadata.

5 Conclusions

The work in this project has shown that already a substantial corpus of metadata is available. The amount and quality of these contextual metadata is highly dependent on the domain, the structure and goals of the RIs (project-centric, service-centric, resource-centric) and the metadata schemas applied in the RIs. A major problem is that quite often the contextual metadata are not explicitly identifiable in dedicated fields of

metadata schemas but are distributed over several fields or implicitly included in the text of other fields. The situation would considerably improve, if the contextual metadata specified in this report could be identified and if possible, the information cleaned and summarised in a separate entity 'research project', to which other information objects related to a 'research project' could refer. It needs to be noted that this approach is possible only to research project oriented RIs and it is out of scope for resource oriented RIs. For legacy data, this can be done retrospectively but, in the future, a prospective approach would be preferable. A further significant step forward could be the introduction of PIDs for 'research projects' as discussed in the FAIRCORE4EOSC project. It should be, however, clear that this approach does not work for all RIs and not all relevant contextual metadata would be part of a 'research project'.

The past decade has been marked by concerns regarding the replicability and reproducibility of published research in different areas of the social sciences and life sciences (e.g., Hensel, 2021; Errington et al., 2021). What is needed is a common generic vocabulary, with which to describe, compare, assess and discuss metadata schemes – in this particular case in terms of the contextual metadata they support. This work could be seen as an attempt to address that problem. Of major importance here is to differentiate between what can be done on a generic level to be really useful and what has to be kept domain specific. The project has explored this issue from the viewpoint of the literature and the participating RIs, has provided an approach for a framework for contextual metadata and has made a proposal on how to implement it via research graphs and other approaches. There is high potential that with a better and more structured approach to contextual metadata a positive effect on replicability and reproducibility of research can be reached. More and deeper discussion on the proposal and how to implement it, is needed as well as extension to other fields and domains to bring this issue forward.

Acknowledgement

The EOSC Future project is co-funded by the European Union Horizon 2020 Framework Programme, Call INFRAEOSC-03-2020, Grant Agreement Number 101017536, Science Project: COVID-19 metadata findability and interoperability in EOSC (META-COVID).

References

Al-Ababneh, M.M. (2020) 'Linking ontology, epistemiology and research methodology', *Science Philosophy*, Vol. 8, pp.75–91.

Batista, D., Gonzalez-Beltran, A., Sansone, S.A. and Rocca-Serra, P. (2022) 'Machine actionable metadata models', *Scientific Data*, Vol. 9. Doi: 10.1038/s41597-022-01707-6.

Becnel, L.B., Hastak, S., Ver Hoef, W., Milius, R.P., Slack, M., Wold, D., Glickman, M.L., Brodsky, B., Jaffe, C., Kush, R. and Helton, E. (2017) 'BRIDG: a domain information model for translational and clinical protocol-driven research', *Journal of the American Medical*, Vol. 24, pp.882–890. Doi: 10.1093/jamia/ocx004.

Benardou, A., Constantopoulos, P., Dallas, C. and Gavrilis, D (2010) *A conceptual model for scholarly research activity*. Available online at: http://hdl.handle.net/2142/14945

Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M. and Schindler, U. (2017) *The Scholix Framework for Interoperability in Data-Literature Information Exchange*, D-Lib Magazine, Vol. 23.

Campos, P.M.C., Reginato, C.C. and Almeida, J.P.A. (2019) 'Towards a core ontology for scientific research activities', in Guizzardi, G., Gailly, F. and Maciel, R.S.P. (Eds): *Advances in Conceptual Modeling*, Springer, Cham. Doi: 10.1007/978-3-030-34146-6_1.

Canham, S. (2023) *ECRIN Metadata Schemas for Clinical Research*, Report. Doi: 10.5281/zenodo.8368709.

Chalmers, I. and Glasziou, P. (2009) 'Avoidable waste in the production and reporting of research evidence', *The Lancet*, Vol. 374, pp.86–89. Doi: 10.1016/S0140-6736(09)60329-9.

David, R., Ohmann, C., Boiten, J.W., Cano Abadia, M., Bietrix, F., Canham, S., Chiusano, M.L., Dastrù, W., Laroquette, A., Longo, D., Mayrhofer, M.T., Panagiotopoulou, M., Richard, S.R., Goryanin, S. and Verde, P.E. (2022) 'An iterative and interdisciplinary categorisation process towards FAIRer digital resources for sensitive life-sciences data', *Scientific Report*, Vol. 12. Doi: 10.1038/s41598-022-25278-z.

Donabedian, A. (1980) *The Definition of Quality and Approaches to Its Assessment, Explorations in Quality Assessment and Monitoring*, Band 1, Health Administration Press.

Errington, T.M., Denis, A., Perfetto, N., Iorns, E. and Nosek, B.A. (2021) 'Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology', *eLife*, Vol. 10. Doi: 10.7554/eLife.67995.

Hensel, P.G. (2021) 'Reproducibility and replicability crisis: how management compares to psychology and economics – a systematic review of literature', *European Management Journal*, Vol. 39, pp.577–594.

Johnson, D., Batista, D., Cochrane, K., Davey, R.P., Etuk, A., Gonzalez-Beltran, A., Haug, K., Izzo, M., Larralde, M., Lawson, T.N., Minotto, A., Moreno, P., Nainala, V.C., O'Donovan, C., Pireddu, L., Roger, P., Shaw, F., Steinbeck, C., Weber, R.J.M., Sansone, S.A. and Rocca-Serra, P. (2021) 'ISA API: an open platform for interoperable life science experimental metadata', *GigaScience*, Vol. 10, No. 9. Doi: 10.1093/gigascience/giab060.

Juul, S., Nielsen, E.E., Feinberg, J., Siddiqui, F., Jørgensen, C.K., Barot, E., Nielsen, N., Bentzer, P., Veroniki, A.A., Thabane, L., Bu, F., Klingenberg, S., Gluud, C. and Jakobsen, J.C. (2020) 'Interventions for treatment of COVID-19: a living systematic review with meta-analyses and trial sequential analyses (The LIVING Project)', *PLoS Medicine*. Doi: 10.1371/journal.pmed.1003293.

Karki, R. (2022) *COVID-19 knowledge graph: a semantic resource embedding biological and chemical entities (v1.0.0)*, Report. Doi: 10.5281/zenodo.7351221.

Karki, R., Gadiya, Y., Zaliani, A. and Gribbon, P. (2023) 'Mpox knowledge graph: a comprehensive representation embedding chemical entities and associated biology of Mpox', *Bioinformatics Advances*. Doi: 10.1093/bioadv/vbad045.

Kleemola, M. (2020) 'SSHOC metadata interoperability aspects', Presented at the *Realising the European Open Science Cloud – Towards a FAIR Research Data Landscape for the Social Sciences, Humanities and Beyond (RealisingEOSC)*, *Event: Realising the European Open Science Cloud – Towards a FAIR Research Data Landscape for the Social Sciences*, Humanities and Beyond. Doi: 10.5281/zenodo.4279855.

Lim, C., Lu, S., Chebotko, A. and Fotouhin, F. (2010) 'Prospective and retrospective provenance collection in scientific workflow environments', IEEE International Conference on Services Computing, Miami, FL, USA, pp.449–456.

Luff, R., Byatt, D. and Martin, D. (2015) *Review of the Typology of Research Methods within the Social Sciences*, National Centre for research Methods Report.

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J. and Principe, P. (2019) *The OpenAIRE Research Graph Data Model (1.3)*, Report. Doi: 10.5281/zenodo.2643199.

Merino-Martinez, R., Norlin, L., Van Enckevort, D., Anton, G., Schuffenhauer, S., Silander, K., Mook, L., Holub, P., Bild, R., Swertz, M. and Litton, J.E. (2016) 'Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 Core), *Biopreserv Biobank*, Vol. 14, pp.298–306. Doi: 10.1089/bio.2015.0070.

Moreau, L., Batlajery, B.V., Huynh D.T., Michaelides, D. and Packer, H.A. (2018) 'Templating system to generate provenance', *IEEE Transactions on Software Engineering*, Vol. 44, pp.103–121.

Ohmann, C., Canham, S. and Panagiotopoulou, M. (2022a) *Protocol of a Qualitative Study to Characterise the Contextual Metadata and Workflows in Selected Research Infrastructures*, Report. Doi: 10.5281/zenodo.7025319.

Ohmann, C., Canham, S. and Panagiotopoulou, M. (2022b) *Interview Guide for a Qualitative Study to Characterise the Contextual Metadata and Workflows in Selected Research Infrastructures*, Report. Doi: 10.5281/zenodo.7025502.

Pearson, H. (2021) *How COVID broke the evidence pipeline*, Nature – News Feature. Available online at: https://www.nature.com/articles/d41586-021-01246-x

Scapicchio, C., Gabelloni, M., Forte, S.M., Alberich, L.C., Faggioni, L., Borgheresi, R., Erba, P., Paiar, F., Bonmati, L.M. and Neri, E. (2021) 'DICOM – MIABIS integration model for biobanks: a use case of the EU PRIMAGE project', *European Radiology Experimental*, Vol. 5, No. 20. Doi: 10.1186/s41747-021-00214-4.

Schröder, M., Staehlke, S., Groth, P., Nebe, J.B., Spors, S. and Krüger, F. (2022) 'Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation', *Journal of Biomedical Semantics*, Vol. 13, No. 4.

Stocker, M., Heger, T., Schweidtmann, A.M., Ćwiek-Kupczyńska, H., Penev, L., Dojchinovski, M., Willighagen, E., Vidal, M-E., Turki, H.A., Balliet, D., Tiddi, I., Kuhn, T., Mietchen, D., Karras, O., Vogt, L., Hellmann, S., Jeschke, J.M., Krajewski, P. and Auer, S. (2022) 'SKG4EOSC – scholarly knowledge graphs for EOSC: establishing a backbone of knowledge graphs for FAIR scholarly information in EOSC', *Research Ideas and Outcomes*, Vol. 8. Doi: 10.3897/rio.8.e83789.

Thiese, M.S. (2014) 'Observational and interventional study design types; an overview', *Biochemia Medica*, Vol. 24, pp.199–210.

Tobi, H. and Kampen J.K. (2018) 'Research design: the methodology for interdisciplinary framework', *Quality and Quantity*, Vol. 52, pp.1209–1225.

Tong, A., Sainsbury, P. and Craig, J. (2007) 'Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups', *International Journal Quality in Health Care*, Vol. 19, pp.349–357. Doi: 10.1093/intqhc/mzm042.

Vasa, C., Fairweather, E., Danger, R. and Corrigan, D. (2017) 'Templates as a method for implementing data provenance in decision support systems', *Journal of Biomedical Informatics*, Vol. 65, pp.1–21.

Wittner, R., Holub, P., Müller, H., Geiger, J., Goble, C., Soiland-Reyes, S., Pireddu, L., Frexia, F., Mascia C., Fairweather, E., Swedlow, J.R., Moore, J., Strambio, C., Grunwald, D. and Nakae, H. (2021) 'ISO 23494: biotechnology – provenance information model for biological specimen and data', in Glavic, B., Braganholo, V. and Koop, D. (Eds): *Provenance and Annotation of Data and Processes*, Springer, Cham. Doi: 10.1007/978-3-030-80960-7_16.

Note

1 Appendices are available on request from authors.