Quality Assessment of Open Datasets Metadata

Šlibar, Barbara

University of Zagreb

FACULTY OF ORGANIZATION AND INFORMATICS

Barbara Šlibar

QUALITY ASSESSMENT OF OPEN DATASETS METADATA

DOCTORAL THESIS

Varaždin, 2024

Sveučilište u Zagrebu

FAKULTET ORGANIZACIJE I INFORMATIKE

Barbara Šlibar

PROCJENA KVALITETE METAPODATAKA OTVORENIH SKUPOVA PODATAKA

DOKTORSKI RAD

Varaždin, 2024.

University of Zagreb

FACULTY OF ORGANIZATION AND INFORMATICS

Barbara Šlibar

QUALITY ASSESSMENT OF OPEN DATASETS METADATA

DOCTORAL THESIS

Supervisors:
Full Prof. Diana Šimić, Ph.D.
Full Prof. Nina Begičević Ređep, Ph.D.

Varaždin, 2024

INFORMATION ABOUT SUPERVISORS

Diana Šimić was born in Zagreb, where she graduated in mathematical informatics and statistics at the Mathematical Department of the Faculty of Science at the University of Zagreb. She obtained her master's degree at the same faculty, and she earned her doctorate at the School of Medicine of the University of Zagreb. She worked as a programmer at RIZ Transmitter Factory, as a researcher, head of the Laboratory for Biomathematics, and a member and president of the Management Board of the Institute for Medical Research and Occupational Health. She was the assistant minister for informatics at the Ministry of Science and Technology and the deputy state secretary in the Central State Office for e-Croatia. She also worked as a consultant for e-government and e-business in Infodom. Since 2010, she has been working at the Faculty of Organization and Informatics of the University of Zagreb (FOI), where in 2021 she was elected to the position of full professor in permanent tenure on subjects related to statistics, quantitative and scientific methods in information sciences. Since 2016, she has been the head of the doctoral study of Information Sciences. She is a member of the Department of Quantitative Methods and the Laboratory for Learning and Academic Analytics. She participated in more than 40 scientific projects, supervised 4 doctoral, 12 graduate and 2 bachelor theses. Her 43 papers are indexed in the Web of Science Core Collection databases (499 citations, h-index 14), and together with other databases on this platform, 60 of her papers are indexed. She is also the co-author of 11 books and 4 chapters in books. She was an invited lecturer at more than 20 conferences. She is a member of the Association for Computing Machinery (ACM), American Statistical Association, and Croatian Biometric Society (where she held the positions of president, vice president, secretary and member of the Presidency and member of the International Program and Organizing Committees of the BIOSTAT Conference). Since 2016, she has been the vice president of the Croatian Informatics Association. She received the Informatics plaque from the Croatian Informatics Association for her contribution to the development of the information society in Croatia (2007), Recognition for a special contribution to the progress and recognisability of the University Computing Centre (2011), Europe's Open Access Champions (2016), Special thanks from the Dean of FOI for contributing to the development of the doctoral study and the development of research methodology (2016), Special thanks from the Dean of FOI for promoting the visibility of the faculty through the e-Schools project (2017), Special recognition from the Dean for exceptional commitment and contribution to the Faculty (2020) and the Lifetime Achievement Award of the Faculty of Organization and Informatics (2023).

Nina Begičević Ređep was born in Varaždin and has become a distinguished figure in the fields of business decision-making and decision theory within the academic community. Presently, she holds the title of full professor at the Faculty of Organization and Informatics of the University of Zagreb (FOI). Her academic journey began at FOI, where she completed her doctoral studies in information science. This foundation set the stage for further specialization at the University of Pittsburgh, Katz Graduate School of Business, facilitated by a scholarship from the U.S. Government. Nina's career at FOI has been marked by significant leadership roles, including her tenure as dean from 2019 to 2023. She has also served as Vice-dean for research and international relations and Vice-dean for business affairs, showcasing her versatility and dedication to the institution. In 2018, she founded the Laboratory for Strategic Planning and Decision Making at FOI. Her expertise in strategic planning and decision making, particularly the multicriteria decision making method known as the analytic hierarchy process, is well-regarded. Nina's teaching responsibilities span undergraduate, graduate, and postgraduate study programmes, reflecting her comprehensive commitment to education. She is recognized for her ability to seamlessly integrate digital technologies into educational contexts, enhancing the learning experience for students across all levels. Nina's scholarly contributions are extensive, with authorship and co-authorship of more than 80 research articles, book chapters, and conference papers. Her research interests encompass a wide range of topics related to business decision-making, strategic planning, innovative and sustainable entrepreneurial solutions, innovative use of digital technologies in education, and the implementation of learning analytics in higher education institutions. She has led and participated in numerous national and international projects, further solidifying her position as a thought leader in her field. In addition to her academic and research endeavors, Nina serves as the President of the Board of Science, International Cooperation, and Art at the University of Zagreb. This role underscores her influence and commitment to advancing scientific research, fostering international collaboration, and promoting the arts within the academic community. Her dedication to her profession and the broader community has not gone unnoticed. Nina has received several recognitions and awards for her scientific work and contributions to society, marking her as a respected figure in her field. Her involvement in editorial boards and as a reviewer for prestigious journals, including the International Journal of the Analytic Hierarchy Process, highlights her commitment to the dissemination of knowledge and the promotion of scientific excellence.

‘Which is more important,’ asked Big Panda, ‘the journey or the destination?’

‘The company,’ said Tiny Dragon.

— James Norbury, *Big Panda and Tiny Dragon*

ACKNOWLEDGMENTS

I am deeply indebted to my supervisors, Diana and Nina, who have supported me with advice and were happy to assist me during this academic journey. Diana, a strong pillar of support, has not only dedicated countless hours to supporting me, but has also provided invaluable insight and has been my go-to person every step of this challenging journey. I am equally grateful to Nina for her generosity in allowing me complete freedom and providing me with the resources I needed. Through her, I had the opportunity to meet well-respected experts from various fields.

I would also like to express my gratitude to my esteemed colleagues — Enrique, who has always challenged me to step out of my comfort zone and take advantage of the opportunities presented to me; Dijana, who was the first to introduce me to the intriguing world of science; Matija, a steadfast ally who was always willing to help and engage in constructive discussions; Darko, a reliable support who always listened to my concerns and skilfully turned them into moments of laughter; Ana and Bogdan, trustworthy individuals with whom I effortlessly established mutual understanding. Also, this experience would not be the same without Jelena, Sabina, Lea, Petra, Marina, Marinko, Jurica, Valentina, Blaženka, Barbi, Tanja, Nikola, Tena, and Marko.

My deepest appreciation goes to my husband Josip, who has been generously sharing his knowledge and experience with me, which has been invaluable on this journey. His unwavering belief in me and his constant encouragement gave me comfort in moments of uncertainty when solutions seemed elusive and the path uncertain. I greatly appreciate his understanding of the time I devote to my research.

A special and heartfelt thanks go to my parents, Zdenka and Ivan, for their unconditional love, care and guidance in raising me without prejudice and shaping me into the person I am today.

I would also like to thank my sister Julija and my brother-in-law Goran. Their understanding of my to-do list and their ability to give me moments of joy provided a much-needed break from my obligations.

Finally, the most special thank you goes to my niece Lea. She has always been considerate of my work, never disturbing me and showing immense pride in my achievements. Lea is an outstanding student, creative, and shares my love for food and music. Her joy in seeing me pursue my PhD has been a constant source of motivation.

ABSTRACT

Open data are an extremely valuable information technology resource for economic, social, and human development, adding new values to the development of society. More and more countries around the world are establishing open data portals at national, regional, and local levels, and the amount of available open data is growing. The usability of open data depends on the quality of their metadata, the evaluation of which is an open research question.

The main objective of the research is to develop a theoretical framework of open metadata quality and to operationalise it through a new composite indicator that enables the comparison of open datasets metadata. The research approach is based on the Methodological framework for design science research and the Methodology for constructing composite indicators, involving predominantly quantitative and, to a minor extent, qualitative research methods. The combination of these two methodologies helped meet the requirements for ensuring scientific contribution in the construction of the composite indicator and for achieving greater practical relevance of the scientific results within the area of information technology and information systems.

The scientific contributions are achieved through the development of a framework and composite indicator, a better understanding of the concept of open (meta)data quality, and empirical research of the public sector metadata quality. This research also contributes to practise. One of the most important practical contributions is that the developed composite indicator for the metadata quality of open datasets can be used for benchmarking purposes.

Keywords: *open government data; metadata; quality; composite indicator; benchmarking*

PROŠIRENI SAŽETAK

Otvoreni podaci izrazito su vrijedan resurs informacijske tehnologije za ekonomski, socijalni i ljudski razvoj koji dodaje nove vrijednosti razvoju društva. Sve više država u svijetu uspostavlja portale otvorenih podataka na nacionalnoj, regionalnoj i lokalnoj razini, čime raste količina dostupnih otvorenih podataka. Iskoristivost otvorenih podataka ovisi o kvaliteti njihovih metapodataka, čije je vrednovanje otvoreno istraživačko pitanje. Stoga je glavni cilj istraživanja razviti teorijski okvir kvalitete otvorenih metapodataka i operacionalizirati ga kroz novi kompozitni indikator koji će omogućiti usporedbu metapodataka otvorenih skupova podataka.

U prvom dijelu ***prvog poglavlja*** predstavljeno je područje istraživanja. Istaknuta je uloga i značaj otvorenih podataka za napredak društva, dan je kratak povijesni pregled ključnih događaja s naglaskom na područje Sjeverne Amerike i Europe te su izdvojeni izazovi i prepreke koji utječu na razvoj otvorenih podataka. Kao jedna od prepreka koje mogu negativno utjecati na uspjeh otvorenih podatka navodi se kvaliteta otvorenih (meta)podataka. Stoga je napravljen pregled postojećih istraživanja koja se bave kvalitetom otvorenih (meta)podataka te su uočeni nedostaci, također istaknuti u sklopu ovog poglavlja.

U drugom dijelu prvog poglavlja, u vezi s identificiranim istraživačkim problemom koji se odnosi na upitnu kvalitetu otvorenih (meta)podataka), definirano je pet ciljeva istraživanja: sintetizirati rezultate prethodnih istraživanja na temu kvalitete otvorenih (meta)podataka i dimenzija identificiranih u svrhu mjerenja istih (C1), definirati teorijski okvir kvalitete metapodataka otvorenih skupova podataka (C2), prikupiti i organizirati podatke o metapodacima s portala otvorenih podataka (C3), definirati kompozitni indikator kvalitete metapodataka otvorenih skupova podataka (C4) te izračunati vrijednosti kompozitnog indikatora na prikupljenim podacima (C5). Uz ciljeve definirana su tri istraživačka pitanja: „Koje su ključne dimenzije kvalitete metapodataka otvorenih podataka?", „Kako mjeriti identificirane dimenzije kvalitete metapodataka?" i „Kako procijeniti kvalitetu metapodataka?". Također, postavljena je hipoteza vezana uz operacionalizaciju teorijskog okvira kvalitete metapodataka otvorenih skupova podataka koja glasi: Razvijeni kompozitni indikator kvalitete metapodataka otvorenih skupova podataka robustan je.

U trećem dijelu prvog poglavlja predstavljen je teorijski i konceptualni okvir istraživanja. Navedene su teorije na kojima se temelji istraživanje, uključujući i metodološke teorije. Istraživanje se temelji na kombinaciji dviju metodologija: metodološkog okvira znanosti o dizajniranju i metodologije za konstruiranje kompozitnih indikatora. Sinergija tih dviju metodologija može pomoći u ispunjavanju zahtjeva za osiguravanjem znanstvenog doprinosa u izradi kompozitnih indikatora, kao i u ispunjavanju zahtjeva za većom praktičnom relevantnošću znanstvenih rezultata u području informacijskih tehnologija i informacijskih sustava. Stoga su aktivnosti metodološkog okvira znanosti o dizajniranju povezane s koracima metodologije za konstruiranje

kompozitnih indikatora.

U ***drugom poglavlju*** prikazani su sustavi za upravljanje podacima koji se koriste za otvorene podatke javnih uprava te je predstavljen povijesni razvoj metapodataka, njihove osnovne komponente, međunarodne norme relevantne za metapodatke te rezultati prethodnih istraživanja kvalitete metapodataka. Također, napravljena je harmonizacija metapodatkovnih polja dvaju relevantnih metapodatkovnih standarda koji definiraju strukturu i semantiku podataka te istovremeno omogućuju bilježenje informacija o skupovima podataka. Rezultati harmonizacije metapodatkovnih polja relevantnih metapodatkovnih standarda (ISO/IEC 11179 i DCAT 2) dostupni su u otvorenom pristupu u repozitoriju otvorenih znanstvenih podataka *Harvard Dataverse*.

U ***trećem poglavlju*** opisana je metodologija rada koja objedinjuje dva metodološka pristupa: metodološki okvir znanosti o dizajniranju i metodologiju za konstruiranje kompozitnih indikatora. Istraživanje je provedeno slijedeći aktivnosti metodološkog okvira znanosti o dizajniranju i odgovarajuće korake metodologije za konstruiranje kompozitnih indikatora. Nadalje, u ovom istraživanju korištene su pretežno kvantitativne, a u manjoj mjeri i kvalitativne istraživačke metode.

U ***četvrtom poglavlju*** prikazani su rezultati istraživanja. U prvom dijelu ovog poglavlja prikazani su rezultati istraživanja teorijskog okvira kvalitete metapodataka koji se oslanja na sustavni pregled literature, analizu i mapiranje metapodatkovnih polja različitih međunarodnih standarda i specifikacija sustava za upravljanje otvorenim podacima te istraživanje mišljenja stručnjaka s ciljem provjere sadržajne valjanosti. Inicijalno razvijen teorijski okvir kvalitete metapodataka otvorenih skupova podataka sadrži 71 individualni indikator, od kojih svaki pripada jednoj od osam dimenzija s obzirom na svojstva metapodataka i jednoj od pet dimenzija s obzirom na svojstva skupova podataka (dostupan je u otvorenom pristupu na repozitoriju *Harvard Dataverse*). Provjerom sadržajne valjanosti razvijenoga teorijskog okvira relevantnima su se pokazala 32 individualna indikatora. Relevantni indikatori raspoređeni su u šest dimenzija s obzirom na svojstva metapodataka te u pet dimenzija s obzirom na svojstva skupova podataka.

U drugom dijelu poglavlja prikazani su koraci u razvoju i validaciji kompozitnog indikatora te rezultati empirijskog istraživanja na uzorku otvorenih podataka preuzetih s dvaju portala otvorenih skupova podataka, koji su bazirani na različitim sustavima za upravljanje metapodacima. Izgradnja kompozitnog indikatora kvalitete metapodataka otvorenih skupova podataka temeljila se na teorijskom okviru, a uključivala je: dohvaćanje metapodataka otvorenih skupova podataka s odabranih portala otvorenih podataka (slučajno odabranih 4820 skupova otvorenih podataka s dvaju portala otvorenih podataka, portala otvorenih podataka EU-a i australskog portala otvorenih vladinih podataka), mapiranje metapodatkovnih polja odabranih portala na metapodatkovna polja navedena u relevantnim metapodatkovnim standardima (pojedina metapodatkovna polja nisu pronađena ni na jednom od odabranih portala), izračunavanje vrijednosti/skorova relevantnih individualnih indikatora, analiziranje izračunatih vrijednosti multivarijatnom analizom, određivanje relativne važnosti odnosno pondera individualnih indikatora i dimenzija analitičkim hijerarhijskim procesom te agregiranje vrijednosti u jednu vrijednost, tzv. skor kompozitnog indikatora, primjenom linearne agregacije. Ponderi individualnih indikatora i dimenzija, izračunati skorovi individualnih indikatora te skorovi (pod)dimenzija i kompozitnog indikatora dostupni su kao otvoreni znanstveni podaci na repozitoriju *Harvard Dataverse*. Kompozitni indikator validiran je evaluacijom robusnosti razvijenoga kompozitnog indikatora, što je uključivalo primjenu analize osjetljivosti i analize nesigurnosti.

U ***petom poglavlju*** najprije su predstavljeni rezultati istraživanja u kontekstu postavljenih ciljeva, istraživačkih pitanja i hipoteza. Tako je u sklopu prve aktivnosti Objašnjenje problema dobiven odgovor na prva dva istraživačka pitanja. Odgovor na istraživačko pitanje „Koje su ključne dimenzije kvalitete metapodataka otvorenih podataka?" glasi: Pet je dimenzija kvalitete metapodataka otvorenih skupova podataka s obzirom na svojstva skupova podataka, a to su pronalažljivost, dohvatljivost, interoperabilnost, ponovna upotrebljivost i kontekstualnost. Osam je dimenzija kvalitete s obzirom na svojstva samih metapodataka, a to su potpunost, usklađenost, koherentnost, točnost, otvorenost, dohvatljivost, razumljivost i pravovremenost. Odgovor na istraživačko pitanja „Kako mjeriti identificirane dimenzije kvalitete metapodataka?" glasi: Identificirane dimenzije kvalitete metapodataka mjerimo pomoću indikatora kvalitete za metapodatke otvorenih skupova podataka, tj. primjenom metrika nad svojstvima/atributima. U sklopu prve aktivnosti ostvaren je i prvi cilj istraživanja (C1). Drugi cilj istraživanja (C2) ostvaren je u aktivnosti Definiranje zahtjeva. Dva cilja (C3 i C4) ostvarena su u sklopu aktivnosti Dizajn i razvoj artefakta. Peti cilj istraživanja (C5) te treće istraživačko pitanje pokriveni su aktivnošću Demonstracija artefakta. Odgovor na istraživačko pitanje „Kako procijeniti kvalitetu metapodataka?" glasi: Kvaliteta metapodataka otvorenih skupova podataka procjenjuje se primjenom kompozitnog indikatora, koji je razvijen na temelju teorijskog okvira kvalitete metapodataka otvorenih skupova podataka, na podacima. Rezultatima aktivnosti Evaluacija artefakta potvrđena je postavljena hipoteza da je razvijeni kompozitni indikator kvalitete metapodataka otvorenih skupova podataka robustan.

Drugi dio petog poglavlja sadrži sažetu usporedbu rezultata provedenog istraživanja s prethodnima, usko povezanima s fokusom i naporima ovog istraživanja, da bi se naglasila važnost i doprinos provedenog istraživanja za područje otvorenih podataka.

Također, peto poglavlje sadrži opis ograničenja provedenog istraživanja, od kojih su neka istaknuta u nastavku. Jedno je od ograničenja da su i procesi i rezultati aktivnosti Objašnjenje problema do određene mjere subjektivni te ovise o znanju i vještinama autorice. Primjerice, proces usklađivanja atributa metapodataka različitih metapodatkovnih standarda podložan je mogućim pogrešnim tumačenjima jer se usklađivanje provodi mapiranjem atributa na temelju njihovog semantičkog značenja. Nadalje, budući da su skripte u programskom jeziku *R* razvijene za analizu podataka uz primjenu odgovarajućih istraživačkih metoda unutar gotovo svih aktivnosti metodološkog okvira, postoji mogućnost da je došlo do defekata tijekom implementacije.

U ***šestom, završnom poglavlju*** navedeni su znanstveni doprinosi istraživanja, a to su: sistematizacija i sinteza dosadašnjeg znanja u domeni kvalitete otvorenih (meta)podataka i dimenzija identificiranih u svrhu mjerenja istih, razvoj teorijskog okvira kvalitete metapodataka otvorenih skupova podataka, razvoj kompozitnog indikatora kvalitete metapodataka otvorenih skupova podataka, rezultati empirijskog istraživanja kvalitete metapodataka otvorenih podataka. Ujedno je detaljnije opisano kako je svaki od prethodno navedenih doprinosa postignut.

Osim znanstvenih doprinosa istraživanja, u završnom poglavlju istaknuti su i praktični doprinosi, a neki od njih navedeni su u nastavku. Razvijen i validiran teorijski okvir obuhvaća dva pogleda na procjenu kvalitete metapodataka otvorenih skupova podataka: prvi, koji više preferira akademska zajednica, usmjeren je na svojstva metapodataka, dok je drugi, koji više preferira praktična zajednica, usmjeren je na svojstva skupova podataka. Nadalje, ustanovljeno je da na vrlo velikom slučajnom uzorku otvorenih skupova podataka nedostaju metapodatkovna polja za pojedine indikatore kvalitete koje stručnjaci smatraju relevantnima. Također, samo manji dio promatranih skupova podataka postigao je višu vrijednost kompozitnog indikatora. Među ostalim doprinosima, razvijeni kompozitni indikator pokazao se kao korisno sredstvo za usporedbu različitih skupova otvorenih podataka i portala koji ih nude.

Poglavlje završava prijedlogom smjernica za buduća istraživanja s ciljem rasta i razvoja otvorenih podataka. To uključuje, među ostalim, implementaciju razvijenoga kompozitnog indikatora kao interaktivne *web* aplikacije koristeći paket *Shiny* programskog jezika *R* te ispitivanje povezanosti razvijenoga kompozitnog indikatora s drugim pokazateljima različitih karakteristika javne uprave (transparentnost i otvorenost, razvijenost e-uprave, uključenost građana, inovacijska sposobnost) i dr.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$Q1$	First/Lower quartile
$Q3$	Third/Upper quartile
$\bar{x}$	Mean
$q5$	$5th$ percentile
$q95$	$95th$ percentile
IQR	Interquartile range
SD	Standard deviation
h^2	Communality
S_i	First-order sensitivity index
S_{Ti}	Total effect sensitivity index

LIST OF ABBREVIATIONS

AACR Anglo-American Cataloguing Rules.

AHP Analytic Hierarchy Process.

AI Artificial Intelligence.

ALA American Library Association.

API Application Programming Interface.

C Contextuality.

CC:DA American Library Association's Committee on Cataloging: Description and Access.

CI Composite Indicator.

CR Consistency Ratio.

CVR Lawshe's Content Validity Ratio.

DCAT Data Catalog Vocabulary.

DCAT 2 Data Catalog Vocabulary - Version 2.

DCAT 3 Data Catalog Vocabulary - Version 3.

DCMI Dublin Core Metadata Initiative.

DMS Data Management System.

DSR Design Science Research.

EC European Commission.

EU European Union.

F Findability.

FOIA Freedom of Information Act.

HTML Hypertext Markup Language.

HTTP Hypertext Transfer Protocol.

I Interoperability.

ICT PSP Information and Communication Technologies Policy Support Program.

IEC International Electrotechnical Commission.

II Individual Indicator.

ISO International Organization for Standardization.

ISO 14721 ISO 14721 Space data and information transfer systems — Open archival information system (OAIS) — Reference model.

ISO 14721:2003 ISO 14721:2003 Space data and information transfer systems — Open archival information system (OAIS) — Reference model.

ISO 14721:2012 ISO 14721:2012 Space data and information transfer systems — Open archival information system (OAIS) — Reference model.

ISO 19115-1:2014 ISO 19115-1:2014 Geographic information — Metadata.

ISO 639 ISO 639 Codes for individual languages and language groups.

ISO 8601 ISO 8601 Date and time — Representations for information interchange.

ISO 9000:2015 ISO 9000:2015 Quality management systems — Fundamentals and vocabulary.

ISO/IEC 11179 ISO/IEC 11179 Information technology — Metadata registries (MDRs).

ISO/IEC 11179-33:2023 ISO/IEC 11179-33:2023 Information technology — Metadata registries (MDRs) — Part 33: Metamodel for data set registration.

ISO/IEC 11179-3:2013 ISO/IEC 11179-3:2013 Information technology — Metadata registries (MDRs) — Part 3: Registry metamodel and basic attributes.

ISO/IEC 11179-6:2013 ISO/IEC 11179-6:2015 Information technology — Metadata registries (MDRs) — Part 6: Registration.

ISO/IEC 11179-7:2019 ISO/IEC 11179-7:2019 Information technology — Metadata registries (MDRs) — Part 7: Metamodel for data set registration.

IT Information Technology.

JCR Joint Research Centre.

KOS Knowledge Organisation System.

MARC MAchine-Readable Cataloging.

MDR Metadata Registry.

NASA National Aeronautics and Space Administration.

OAIS Open Archival Information System.

OD Open Data.

OECD Organisation for Economic Co-operation and Development.

OGD Open Government Data.

PC Principal Component.

PCA Principal Components Analysis.

PSI Public Sector Information.

R Retrievability.

SA Sensitivity Analysis.

Share-PSI 2.0 Shared Standards for Open Data and Public Sector Information.

SLR Systematic Literature Review.

TODO Twinning Open Data Operational.

U Re**U**sability.

UA Uncertainty Analysis.

URI Uniform Resource Identifier.

URL Uniform Resource Locator.

W3C World Wide Web Consortium.

GLOSSARY

Accuracy assesses the precision of the information contained in the metadata. It can also be defined as the correspondence of metadata to actual data (more specifically, to the resources) or to quality certification document or similar documents (Šlibar, Oreški, & Begičević Ređep, 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier, Umbrich, & Polleres, 2016).

Application programming interface (API) is an important intermediary that enables smooth communication and data exchange between websites and software or applications. It is an important feature of data management systems (Ali, Alexopoulos, & Charalabidis, 2022; Braunschweig, Eberius, Thiele, & Lehner, 2012).

Artefact is an object created by a human for the purpose of solving a practical problem (Johannesson & Perjons, 2014).

Coherence measures the degree to which all metadata uniformly describe a particular object (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016).

Completeness refers to the extent of the information present in the metadata (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016).

Composite indicator (CI), which is based on an underlying model, combines multiple separate indicators into a single index to assess a multidimensional notion that cannot be measured with a single indicator alone (OECD, EU, & JCR, 2008).

Conformance means the absence of contradictions and reflects the logical consistency of the metadata with its preceding values, established norms, standards and other relevant criteria (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016).

Content validity refers to the extent to which a measure (i.e., the artefact developed) captures all aspects of a given concept (e.g., the quality of metadata for open datasets) (Allen, 2017b).

Contextuality is the extent to which the user can obtain additional information about the data (e.g., origin, quality, copyright statements, date of publication) (Consortium of data.europa.eu, n.d.).

Criterion validity refers to the extent to which the results of a new measure (i.e., the artefact developed) agree with the results of another measure that is already considered valid, i.e., the criterion variable (Allen, 2017a).

Data catalogue serves as a centralized metadata repository that organisations rely on for the purpose of data management (Herzberg, 2022).

Data management refers to the systematic development, implementation and monitoring of policies, measures, protocols and strategies designed to deliver, regulate, protect, and improve the value of data and information throughout its lifecycle (DAMA International, 2017).

Data management system (DMS) refers to specialised software or a collection of tools structured to streamline and optimise the management of data (Watson, 2016).

Data repository / archive represents a data storage entity where data is partitioned for an analytical or reporting purposes. It is a form of sustainable information infrastructure since it offers data storage and access over a long period of time (Fitzgerald, 2022; U.S. Geological Survey, n.d.).

Dataset 'a collection of data published or curated by a single agent, and available for access or download in one or more representations' (W3C, 2020). The term 'dataset' within this research is used both for a class and for its instantiation.

Design science is an epistemological paradigm for carrying out research, i.e., it leads research toward artifact design and problem solving (Dresch, Lacerda, & Antunes Júnior, 2014).

Design science research (DSR) is a research method that operationalizes the research whose objective is to solve a problem by creating an artefact, or rather, something new under the paradigm of design science (Dresch et al., 2014; Hevner, March, Park, & Ram, 2004).

Dimension represents a hierarchical level of analysis that is an outcome of an aggregation function. It describes the scope of objectives, indicators, and variables (OECD et al., 2008).

Element set conveys the structure and semantics of a collection of elements. Sometimes it can denote a metadata schema or a data dictionary. One of widely used is the Dublin Core Metadata Element Set (DCMI Usage Board, 2012; Zeng & Qin, 2022).

Findability is the extent to which humans and machines can easily discover (meta)data. In addition to information that helps both humans and machines to identify data uniquely and unambiguously, information about the temporal and geographic area(s) covered by the data is also relevant for this dimension (Consortium of data.europa.eu, n.d.; Deutz et al., 2020).

First-order sensitivity index denotes the main effect contribution of each input uncertainty to the variance of the output (OECD et al., 2008; Saltelli et al., 2008).

Free - for a dataset - refers to two aspects of accessing datasets, and those are: the possibility of access and the cost of access. Free access means that any person should be able to access the data at any time without revealing their identity. If there is a cost to access open data, then the cost should be negligible. It should be clear that the cost refers to what an end user spends to access the data, not what the public or private data publisher spends (Šlibar, Oreški, & Kliček, 2018).

Horn's method is a simulation-based method to determine how many factors / principal components should be retained (Horn, 1965).

Individual indicator represents the result of application of a metrics over a property/attribute (i.e., metadata statement) (OECD et al., 2008).

Interoperability is the extent to which different applications and systems can successfully communicate and exchange data with unambiguous, shared meaning. Interoperability implies both syntactic interoperability (compatible formats and protocols) and semantic interoperability (uniform codification of data) (Consortium of data.europa.eu, n.d.; Deutz et al., 2020).

Loadings are the elements of the eigenvectors that represent the weights of each variable on the factor / principal component (Field, Miles, & Field, 2012).

Machine-readable means that the data is stored in a data format that can be read and processed automatically by a machine or computer, such as CSV, JSON, XML, etc. It also means that the data is structured (Šlibar et al., 2018).

Metadata are data about data (Zeng & Qin, 2022; Riley, 2017).

Metadata description is a collection of one or more statements about a single entity that may contain a uniform resource identifier associated with the described entity (Zeng & Qin, 2022; Powell, Nilsson, Naeve, Johnston, & Baker, 2007).

Metadata quality, just like the quality of data or quality in general, is a multidimensional concept of 'fitness for use' or, more recently, 'fitness for a particular purpose' (Attard, Orlandi, Scerri, & Auer, 2015; Kučera, Chlapek, & Nečaský, 2013; Wand & Wang, 1996; Juran & DeFeo, 2010; Šlibar & Mu, 2022; Tauberer, 2012).

Metadata record, also known as a description set, is a compilation of one or more metadata descriptions, each consisting of one or more metadata statements, relating to a single entity (Zeng & Qin, 2022).

Metadata statement is a fundamental component of metadata that represents a property in conjunction with its associated value. An interchangeable term for a metadata statement is 'metadata field' (Zeng & Qin, 2022; Powell et al., 2007).

Metric, in a pragmatic artefact classification, is an artefact type which can be defined as a mathematical model that can be used to quantify various aspects of systems or methods (Dresch et al., 2014; Johannesson & Perjons, 2014; Offermann, Blom, Schönherr, & Bub, 2010). At the level of the individual indicators, the metrics represent the quantification of metadata quality properties.

Non-proprietary file format is one that is not owned and controlled by a company. Data in this format does not require proprietary software to be reliably read. An example of such a file format is Comma Separated Value, as opposed to Microsoft Excel Spreadsheet, which is proprietary (Šlibar et al., 2018).

Open data (OD) are data that can be freely used, reused and redistributed by anyone, with the highest standards of provenance and openness. It should also be publicly available via a public server and published in a machine-readable, preferably non-proprietary format (Šlibar et al., 2018; Attard et al., 2015; Pasquetto, Randles, & Borgman, 2017).

Open data portal is a web platform primarily designed to serve as a data catalogue, facilitating the publication of data. Usually, it is powered by a data management system (Neumaier et al., 2016; Šlibar & Mu, 2022; Milić, Veljković, & Stoimenov, 2018).

Open government data (OGD) are data collected by the public sector within its jurisdiction that can be FREELY reused for ANY purpose (Varga & Vračić, 2015; Musa, Bebić, & Đurman, 2015).

Openness refers to the extent to which data/content (e.g., metadata, resources) conforms to open licences, are non- proprietary, and are machine-readable (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016).

Public sector includes public administration, public services and enterprises that are majority-owned by the state and/or local and regional self-governing units (Republic of Croatia Ministry of Justice and Public Administration, 2015; EC & Directorate-General for Employment, Social Affairs and Inclusion of the EC, 2010).

Publicly available means that data should be published in such a way that access to the data is granted without unfair restrictions or access costs regarding data use. When a resource is published as open data, it is automatically considered publicly available, but not vice versa (Šlibar et al., 2018).

Redistribution refers to an action where data is shared differently than before when it is published along with permission to redistribute it. It also covers combining multiple datasets together (Šlibar et al., 2018).

Relevance refers to a business problem. Any design science research effort should be aligned with the practitioners' community. Criteria for assessing relevance focus on representational fidelity and implementability of the artefact (Dresch et al., 2014; Hevner et al., 2004).

Retrievability is the extent to which humans and machines can fetch (meta)data successfully (Consortium of data.europa.eu, n.d.; Deutz et al., 2020; Šlibar et al., 2021).

Reusability is the extent to which (meta)data are well-described so that data can be replicated by different teams within different experimental setups. The information about the terms and conditions on how the data could be accessed and reused is relevant for this dimension, as well as provenance information related to the data creation process, and who can be contacted for more information about the data (Consortium of data.europa.eu, n.d.; Deutz et al., 2020; Association for Computing Machinery, 2020).

Reuse is an action that is permitted when data, once published, can be used again, usually for a purpose other than that for which they were originally collected, and their reuse can only be determined under an open licence by the data-holder (Šlibar et al., 2018).

Rigor refers to the way of conducting research work, especially methodology. Criteria for assessing rigor in design science research focus on applicability and generalisability of the artefact (Dresch et al., 2014; Hevner et al., 2004).

Schema is a machine-readable specification that describes the structure, encoding syntax, rules and formats for a specific set of metadata elements in a formal schema language (Zeng & Qin, 2022).

Sensitivity analysis (SA) is used to estimate which of the input uncertainties contribute to the output uncertainty and by how much. Through the application of sensitivity analysis, it can be recognised which input uncertainties have the most substantial influence on the model, i.e., the composite indicator (thus possibly deserving additional attention), distinguishing them from those that do not (Saisana, Saltelli, & Tarantola, 2005; OECD et al., 2008).

Timeliness refers to the clarity and comprehensibility of the information contained in the metadata (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016).

Total effect sensitivity index represents the overall contribution to the variation in output caused by an input uncertainty and includes both its first-order effect and all higher-order effects resulting from interactions (OECD et al., 2008; Saltelli et al., 2008).

Uncertanty analysis (UA) is referred to as the process by which uncertainty in the inputs spreads through the model/system (here, the structure of the composite indicator) and affects the uncertainty in the outputs (here, the values of the composite indicator). Input uncertainties (assumptions, parameters) in the context of this research are referred to as actions taken during certain steps of constructing a composite indicator, for example, the inclusion or exclusion of individual indicators, the use of alternative methods for data normalisation, and the application of different weighting methods (Saisana et al., 2005; OECD et al., 2008).

Understandability refers to the clarity and comprehensibility of the information contained in the metadata (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016).

Vocabulary comprises terms that are intended for a specific purpose. When speaking about metadata, there are usually two categories of vocabularies: metadata vocabularies, which define data structures, and knowledge organisation systems vocabularies, which serve as value vocabularies (Zeng & Qin, 2022).

CHAPTER 1

INTRODUCTION

In the first chapter, the research topic is introduced (subchapter 1.1), the research objectives, research questions, and a hypothesis are defined (subchapter 1.2), and the theoretical and conceptual frameworks of the research are presented (subchapter 1.3).

In recent years, interest in open data (OD) has grown as they represent an innovation potential for various stakeholders (individuals, private organisations, public organisations), and their use can create added value. The motivations for the widespread adoption of open government data (OGD) vary from country to country and are influenced by economic, cultural, and political considerations. However, a common belief in the potential benefits of open data is something that links these different motivations. The potential benefits have been extensively explored in the academic literature (Kučera & Chlapek, 2014; Attard et al., 2015; Lourenço, 2015; Jethani & Leorke, 2021), with a prominent discussion by Janssen, Charalabidis, and Zuiderwijk (2012), who categorize the benefits of open data into three overarching groups: political and social (e.g., increasing transparency, improving accountability, enhancing participation), economic benefits (e.g., creating new products and services, fostering innovation), and operational and technical benefits (e.g., facilitating access to data, facilitating discovery of data, enabling the blending of public and private data). These positive impacts are also frequently mentioned in official documents and press releases about open data from various governmental and non-governmental organisations (Publications Office of the European Union, 2020; US General Services Administration, US Office of Government and Information Services, & US Office of Management and Budget, n.d.; Open Knowledge Foundation, n.d.). Additionally, benefits of open data are also briefly discussed in the author's review paper (Šlibar et al., 2021). As elaborated by (Verhulst & Young, 2017; Huyer, van Knippenberg, & Publications Office of the European Union, 2020), there is a lot of research focusing on the potential of open data to produce positive outcomes, while real evidence is still lacking. However, there are studies that examine the actual impact of open data based on existing case studies, reports, indices, and data (Verhulst & Young, 2017; Huyer et al., 2020; Davies, Walker, Rubinstein, & Perini, 2019). Moreover, guidelines on how to analyse and capture evidence of what contributes to successful open data efforts can be found in (Verhulst & Young, 2017; Young, Zahuranec, Verhulst, & Gazaryan, 2021).

The open data movement began with the idea of opening up government data, as described below. It should be emphasised that the following brief history shows important milestones and the development of open data in relation to the two regions: North America and Europe. The origins of the open data movement can be traced back to the late 20th century, when governments began to recognise the importance of making public sector information available to the public.

In the United States, this movement took shape with the passage of the Freedom of Information Act (FOIA) in 1966 (Davies et al., 2019). However, these early efforts encountered limitations due to a lack of digital infrastructure. With the rapid advancement of information technology, crucial events for the development of the open data movement began to occur in the early 21st century. The European Union's Public Sector Information (PSI) Directive[1], introduced in 2003 and amended in 2013, played a crucial role in advancing open data (Directorate-General for Communications Networks, Content and Technology of the EC & EC, n.d.-a). This directive provided a legal framework for governing the release of public sector information for reuse. It laid the foundation for the practise of open data in the European Union (EU). In 2007, a significant event took place in Sebastopol, California, where like-minded people met to discuss the emerging concept of open data (Davies et al., 2019; Data.gov admins, 2013). This conference marked a pivotal moment in the movement's history, bringing together pioneers who recognised the potential of sharing government data with the public. Consequently, in 2009, the United States Administration of Barack H. Obama issued the Memorandum on Transparency and Open Government, which emphasised the principles of transparency, participation and collaboration and led to the launch of Data.gov, a central place to access open government data (Davies et al., 2019; Data.gov admins, 2013). Building on the PSI Directive, the European Commission (EC) adopted the new Directive on Open Data and Re-use of Public Sector Information[2], known as the Open Data Directive, which came into force on 16 July 2019 to remove the remaining barriers to the reuse of publicly funded information and update the legal framework to reflect changes in digital technology (Directorate-General for Communications Networks, Content and Technology of the EC & EC, n.d.-a, n.d.-b). Moreover, Davies et al. (2019) discussed in detail how and to what extent countries and regions around the world have started to embrace and develop open data. In addition, international forums and organisations such as the Open Government Partnership [3], the Open Data Charter[4], the World Wide Web Foundation[5], the Open Knowledge Foundation[6], The Governance Lab[7] and many others should not be forgotten as they play a key role in promoting open data practises on a global scale.

In parallel with the development of open data, numerous obstacles have emerged that both the practical and scientific communities are trying to solve or at least minimise. Although many studies examine potential barriers and their impact on open data initiatives, only a few are described in more detail below. Davies et al. (2019) divide the issues affecting open data into seven broad groups:

- Algorithms and artificial intelligence (AI), e.g., OD communities' slow engagement with

[1] https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32003L0098
[2] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32019L1024
[3] https://www.opengovpartnership.org/
[4] https://opendatacharter.net/
[5] https://webfoundation.org/
[6] https://okfn.org/en/
[7] https://thegovlab.org/

AI, lack of approaches to address opacity of AI algorithms;

- Data infrastructure, e.g., varying quality of infrastructure across countries, lack of investment in data infrastructure building blocks, greater consideration of ethical principles in data collection, management, and use;
- Data literacy, e.g., more open data is made available than actually used, lack of systematic efforts to improve data literacy;
- Gender equity, e.g., gender bias in data collection and sharing, and exclusionary patterns;
- Indigenous data sovereignty, e.g., concerns about the rights of individuals and organisations to manage data pertaining to them, ownership issues;
- Measurement, e.g., overlap with similar efforts to measure open data readiness, openness, implementation, etc., measurement results are primarily used for one-time reports rather than ongoing research;
- Privacy, e.g., identify as many potential issues as possible, challenge how to balance transparency and privacy.

Berends, Carrara, Vollers, and Publications Office of the European Union (2020) examined the challenges faced by data providers and data users in their work with open data. They divided them into six categories (Berends, Carrara, Vollers, & Publications Office of the European Union, 2020):

- Lack of political will or awareness at all levels of government and the political structure of a country are listed as the most important political obstacles.
- Organisational barriers refer to the institutionalisation of open data in public bodies and companies, the skills to work with open data, and the interaction between different actors in the field of open data.
- Financial obstacles are associated with the lack of clear documentation on the advantages of offering open data at no cost, complicating the process for administrations to rationalize revenue loss or, in a wider sense, to grasp the initial value of data publication. Other financial challenges include the lack of funding for various activities related to open data initiatives and the possibility of losing funding due to changes in government priorities.
- Many data publishers and potential users are unaware of the value and potential benefits of open data. This lack of awareness can result in publishing strategies that do not meet users' needs and in potential data users being unaware that certain open datasets are available.
- Legal barriers are reflected in a country's legal framework, which may be unclear, unspecific, or not developed at all; in various privacy restrictions that prevent data from being published; and in the application of an appropriate licence to a dataset.

- Finally, the technical barriers are reflected in the current quality of open data, which is perceived as low, the interoperability of open data portals, the quality of metadata, the lack of standardisation of the various features of open data, etc.

Furthermore, perspectives on barriers that could negatively impact the continued reuse of open data have been collected from organizations working with open data, located in 21 countries, six of which are non-European (Berends, Carrara, Engbers, Vollers, & Publications Office of the European Union, 2020). The most frequently mentioned obstacles were poor quality of open (meta)data[8], poor discoverability due to lack of the right information (i.e., metadata), and lack of standardisation (Berends, Carrara, Engbers, et al., 2020). The recent study by Gao, Janssen, and Zhang (2021) listed several challenges and issues that researchers have faced in the evaluation phases of open government data research.

1.1 Research topic

Along with the growth of the open data movement, numerous impediments have emerged that can negatively affect the success of open data as above-mentioned (Zuiderwijk, Janssen, Choenni, Meijer, & Sheikh Alibaks, 2012; Berends, Carrara, Vollers, & Publications Office of the European Union, 2020; Davies et al., 2019). The questionable quality of open (meta)data is undoubtedly one of them (Neumaier et al., 2016; Reiche, Höfig, & Schieferdecker, 2014; Kučera et al., 2013; Zhang & Xiao, 2020; J. Wang et al., 2023). In the context of this study, the term resource simultaneously refers to an asset that can be used for various purposes, such as data, money, human knowledge and skills, time, but it also refers to the representation of the data itself, which is also called distribution. Although the open data movement emphasizes the availability of data (resources), this does not imply that those on the demand side will search, discover, or use available resources (Neumaier et al., 2016; Zuiderwijk et al., 2012; Kubler, Jérérmy, Neumaier, Umbrich, & Le Traon, 2018; Reiche & Höfig, 2013). So, they need data that will enable them to do this. These data are called metadata. While Zeng and Qin (2022) have provided a detailed overview of existing definitions, the one they believe provides a complete connotation is the definition provided by the American Library Association (ALA) Committee on Cataloging: Description and Access (CC:DA) Task Force on Metadata, which defines metadata as structured data describing the characteristics of entities that are carriers of information and thus enable entity identification, discovery, evaluation, and management functions. It should also be mentioned that the quality of metadata is reflected in the searching, discovering, and usability of resources (Kubler et al., 2018; Reiche & Höfig, 2013).

Therefore, it is essential that one stresses the importance of metadata and its quality. Metadata play an important part in the interaction between user and data repository, which directly affects a discovery system's usability and utility as well as user satisfaction with the outcomes of the information search process. Poor or inconsistent metadata quality can lead to low recall, poor

[8] (Meta)data refers to both data/resources and metadata.

precision, or failure in resource retrieval; low usage of the open data portal and data repository; negative impression of the portal and repository; inefficiency of time and money for all parties involved, both supply and demand, etc (Király, 2019; Palavitsinis, 2014; Zeng & Qin, 2022). Thus, metadata must be of high quality.

To improve data quality in general, it is essential to understand the meaning of data quality. As the concept of quality is cross-disciplinary, there is no generally accepted definition of quality. There is also no consensus on defining data or metadata quality (Bruce & Hillmann, 2004; Tauberer, 2012; Zeng & Qin, 2022; Ochoa & Duval, 2009). However, data quality is perceived as a multidimensional concept of 'fitness for use' or the more recent 'suitability for a particular purpose' (Attard et al., 2015; Kučera et al., 2013; Wand & Wang, 1996; Tauberer, 2012; Šlibar & Mu, 2022; Juran & DeFeo, 2010). As mentioned in the literature, metadata quality is a multidimensional concept, just like data quality. Therefore, it is necessary to identify the quality dimensions as well as the associated indicators for its assessment. Dimension is a key component in quality assessment of (meta)data according to author's analysis (Šlibar et al., 2021). Thus, quality dimension represents an aspect of the (meta)data quality that should be observed during the assessment of data, like open data (Umbrich, Neumaier, & Polleres, 2015; Veljković, Bogdanović-Dinić, & Stoimenov, 2014; Publications Office of the European Union, 2021). Because metadata is data about data, many metadata quality dimensions are the same as data quality dimensions in general, but with slightly different definitions due to metadata characteristics (Riley, 2017; Dekkers, Loutas, De Keyzer, & Goedertier, 2012).

Zeng and Qin (2022) explain how the quality of metadata can be measured at different levels of detail, mostly with the aim of evaluating metadata processes or metadata itself (see subchapter 2.2, page 21). In addition, the evaluation of metadata is based on manual or automated processes. However, automated assessment should not completely exclude human involvement. While manual assessment of metadata quality is commonly used in digital repositories, Ochoa and Duval (2009) point out significant drawbacks of this approach:

- The validity of the manual quality assessment is limited to the time of sampling; if a significant number of new resources are added to the repository, the assessment may lose accuracy, requiring reassessment;

- This approach only allows the inference of average quality; the quality of individual metadata instances can only be determined for those instances that are included in the sample;

- Obtaining quality assessments in this way is expensive, as human experts must assess an increasing number of resources.

Furthermore, according to Reiche and Höfig (2013), a manual evaluation performed by experts is the best way to assess the quality of metadata. Similar to Ochoa and Duval (2009), Reiche and Höfig (2013) emphasize that there are number of obstacles arising from this method of measurement. Some of these include how to assess the quality of an enormous amount of metadata, as part of the metadata is generated automatically due to the interoperability of the repository, metadata need to be evaluated each time a data change occurs, the quantity of data grows with new publishers, etc. Accordingly, manual evaluation is not scalable due to limited resources. Moreover, Kubler et al. (2018) examine the advantages and disadvantages of both approaches. They argue that, while the automated approach provides regular quality reports (e.g., daily or weekly), it has difficulties in seamlessly integrating human insights into a given digital repository. In addition, reviewing the data requires significant resources, especially given the ever-growing volumes of data, which is an advantage of the automated approach. The manual approach, on the other hand, allows the integration of human expertise for in-depth analysis, but this process is time-consuming and cannot be performed frequently.

The literature review, conducted by the author, showed that there is a growing number of studies dealing with the quality of (meta)data in the field of open data (Šlibar et al., 2018, 2021). Also, it is noticed that there is still much room for improvement (Šlibar et al., 2018, 2021). The author identified shortcomings in a systematic literature review and provided guidelines for future research on the quality of open (meta)data (Šlibar et al., 2021). In existing studies, it is often challenging to recognise the level of granularity when assessing quality of open (meta)data, and there is often a mixing of levels, with researchers simultaneously including indicators that relate to more than one level (Šlibar et al., 2021).

Also, a shortcoming is noticed in Kubler et al. (2018) and Vetrò et al. (2016) and it refers to the objectivity of the developed quality assessment frameworks of open data. Namely, Kubler et al. (2018) and Vetrò et al. (2016) used questionnaires (only the survey results or expert judgements were used at multiple stages of framework development) to develop and evaluate the frameworks. In contrast to Vetrò et al. (2016), Kubler et al. (2018) have developed a tool that automatically assesses quality. The benchmarking framework for evaluating the quality of open data portals proposed by Máchová and Lněnička (2017) is akin to that of Vetrò et al. (2016), it was developed and evaluated on the basis of a literature review and a questionnaire, and the assessment is based on a prepared questionnaire. A similar approach is also used in the Global open data index[9], Open Data Inventory[10], Open Data Barometer[11].

Even though open data were already being discussed as early as the 1960s in the US and the 1990s in the EU, the actual availability of open government data is a relatively new development. As a result, the harmonisation of technical terminology in the field of open data has not yet been fully achieved, resulting in a mixing of basic concepts. For example, the synonyms used in the literature for a data management system (DMS) are: platform (V. Wang & Shepherd, 2020), open data management system (Máchová, Hub, & Lněnička, 2018), portal software framework (Neumaier et al., 2016), a framework for publishing open data (Umbrich et al., 2015), etc. Be-

[9] https://index.okfn.org/place/

[10] https://odin.opendatawatch.com/Report/biennialReport2022#sec7

[11] https://opendatabarometer.org/?_year=2017&indicator=ODB

sides, in some existing studies (Vetrò et al., 2016; Schauppenlehner & Muhar, 2018), the term platform is used simultaneously for data management systems and portals, which has a direct impact on the understanding and clarity of the basic concepts in the field. The same quality dimensions have different names in distinct research, which directly affects the understanding and solving of quality problems. Neumaier et al. (2016) include the existence dimension in the framework, and Vetrò et al. (2016) call the same dimension completeness in their paper. Some researchers also use the term quality dimensions, some quality characteristics and some even the term quality metrics, as described in detail by the author in the systematic literature review (Šlibar et al., 2021). The inconsistency of terminology when assessing the quality of metadata in digital repositories is also discussed by Ochoa and Duval (2009). In addition, Nikiforova, Bičevskis, Bičevska, and Odītis (2020) note that, while many existing research papers propose the same name for different data quality dimensions, their meanings are assumed to be different, and vice versa - different names are used to describe the same semantics. Therefore, the terms are used unambiguously in this research (if not, this is clearly indicated) and are defined in order to avoid the transfer of unintended meanings.

Regardless of the research objective and the object of the study, researchers should be concerned with the results of scientific research, i.e., their generalisation. Furthermore, when developing quality assessment frameworks, researchers should think about extensibility, i.e., that these frameworks can be easily extended without major changes to their core structures. Therefore, open data portals should be monitored independently of the data management system on which they are based (e.g., CKAN, DKAN), unless the aim is to specifically address the differences depending on the data management system used. In this way, a developed quality assessment framework can be applied to many portals rather than just a few, giving a more complete picture of the current state of metadata quality of open datasets. A similar approach has been observed in some existing studies (Neumaier et al., 2016; Máchová & Lněnička, 2017; Kubler et al., 2018; Quarati, 2021).

The disadvantages of existing research were identified (insufficient research of open (meta)-data quality, observation of only some dimensions of the open (meta)data quality, the subjectivity of developed frameworks, dependence of developed quality frameworks on a data management system, inconsistency of technical terminology in the field of open data) by the initial literature review on the topic of the quality assessment of open (meta)data and they became a research problem (questionable quality of open (meta)data). The benefit of the partial or complete elimination of identified disadvantages is key to further growth and open data development. Because of the noticed disadvantages in the existing frameworks, there is a need for the development of a metadata quality assessment framework for open datasets. However, it should also be noted that, while this research focuses on open government data, there is no substantial difference between the metadata of open datasets and non-open datasets. In fact, it can be expected that the results of the research on the quality of metadata in open datasets could also be applied in closed systems. However, within the scope of this research, obtaining empirical data to validate the methodology on non-open datasets would be very challenging.

1.2 Research objectives, research questions, and hypothesis

The main objective of this research is to develop a theoretical framework of open metadata quality and operationalize it through a new composite indicator (CI) that will enable the comparison of open datasets metadata. The main purpose of the developed theoretical framework and the corresponding composite indicator is to provide an insight into the metadata quality of open datasets to different open government data stakeholders since searchability, usability, and reusability of data depend on it. According to the identified research problem, the following research objectives (ROs) are proposed:

RO1: To synthesize the results of previous research on the subject of quality of open (meta)data and dimensions identified for the purpose of measuring them.

RO2: To define a theoretical framework of metadata quality for open datasets.

RO3: To collect and organize data about metadata from open data portals.

RO4: To define the composite indicator of metadata quality of open datasets.

RO5: To compute the values of the composite indicator on the collected data.

In addition to the objectives, research is guided by several research questions (RQs):

RQ1: What are the key quality dimensions of open data metadata?

RQ2: How to measure identified metadata quality dimensions?

RQ3: How to assess the metadata quality?

Furthermore, the author posed a hypothesis (H) which is related to the operationalization of the theoretical framework of metadata quality for open datasets, and it is defined as follows:

H1: The developed metadata quality composite indicator for open datasets is robust.

1.3 Theoretical and conceptual framework

The widespread diffusion of design science in other disciplines like engineering and medicine emphasizes the usefulness of using such a scientific approach in conducting studies in the information systems when the research objective is to design, construct or create something new or when the research objective is to solve a problem. One of the first and most eminent discussions of the relation between design and science is one from Herbert Simon's book The Science of the Artificial (Simon, 1996; Dresch et al., 2014). In his book, Simon (1996) separates something that is natural from something that is artificial, or rather, made by a human. Since the science of the artificial is focused on solving problems or creating something new, and not on exploring, describing, explaining, or predicting, the natural and social sciences are not sufficient and there is a need for a design science as a new paradigm for conducting research (Simon, 1996; van Aken, 2004; March & Smith, 1995). Moreover, Simon (1996) discussed how traditional science cannot produce knowledge for things that don't exist. In addition, it is emphasized that there is a need for interaction between the object and the observer during the research process in order to produce genuine knowledge, as mere observation of the study object is not sufficient (Dresch et al., 2014). Another deficiency of traditional science listed in the literature is that the traditional science paradigm cannot overcome the gap between theory and practice because of its exploratory and analytical nature (Romme, 2003; van Aken, 2004). It is stated that the research conducted under the traditional scientific paradigm lacks practical relevance (Romme, 2003; Dresch et al., 2014).

Despite successful efforts to define design science as a legitimate paradigm (Hevner et al., 2004; March & Smith, 1995; Nunamaker, Chen, & Purdin, 1990), the dissemination of design science as an accepted way of thinking or acting in information systems discipline is very slow (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007; Peffers, Tuunanen, & Niehaves, 2018). The possible reasons why design science is being accepted in the information systems discipline at a slow pace are that researchers have difficulties in publishing design science research (DSR) results in highly respected information systems journals, the emphasis on practical contribution is low in relevant information systems journals, editors are primarily focused on scientific contribution, etc (Peffers et al., 2018; Baskerville, Lyytinen, Sambamurthy, & Straub, 2011; Conboy, Fitzgerald, & Mathiassen, 2012). On the other hand, previous studies in the information systems discipline indicated the lack of relevance of the scientific results for practitioners (Peffers et al., 2007, 2018; Österle et al., 2010).

Also, it is necessary to distinguish design science from design science research. While the design science represents the foundation for the exploration of something that is artificial, the design science research represents its operationalization. The design science research operationalizes concepts of the design science paradigm (i.e., definition of design science, artefact, satisfactory solutions, classes of problems, and pragmatic validity) by relevance and rigor principles (Dresch et al., 2014; Hevner et al., 2004).

Since the main objective of this research is to develop a theoretical framework of open metadata quality and operationalize it through a new composite indicator for solving problems related to the quality assessment of open metadata, the approach to this research is shaped by the worldview of the design science which concepts are operationalized by design science research. By looking at worldviews that are often discussed in the literature, this research contains elements of constructivism and pragmatism since author's belief is that there is no single reality or truth (i.e., artefact could be designed and developed in different ways and different stakeholders could have different preferences), the best method is one that solves the problem, research always occurs in social, political or some other contexts, etc (Creswell, 2014;

Farrow, Iniesto, Weller, & Pitt, 2020).

In this regard, methodologically, the research itself is based on the combination of two methodologies, namely, the method framework for design science research proposed by Johannesson and Perjons (2014) and the methodology for constructing composite indicators proposed by Organisation for Economic Co-operation and Development (OECD), EU, and Joint Research Centre (JCR) of the EC (OECD et al., 2008). The question that arises here is why these two methodologies. There are two main reasons. Firstly, the philosophical ideas are not explicitly brought up within the methodology for constructing composite indicators. Furthermore, the construction of a composite indicator can be done routinely. Hence, questions about its scientific value can be raised. Indeed, any developed composite indicator obviously has a practical contribution if it is properly constructed, but does it also have a scientific one? Secondly, there is lack of scientific results that fulfil the DSR's relevance criteria. Even if the research is carried out under the design science and design science research, are the results relevant enough for practitioners?

In general, the combination of these two methodologies can help to meet the requirements of ensuring the scientific contribution of the construction of the composite indicators as well as to meet the requirements for greater practical relevance of scientific results within the area of IT and information systems.

The method framework for design science research under the paradigm of design science can be seen as an outer cycle of conducting the research. It is used to structure the research through logically related activities and guarantee the quality of the research results. The method framework contains five main activities. The objective of each activity can be shortly described as follows (Johannesson & Perjons, 2014):

- Explicate problem - to phrase the problem exactly and justify its relevance to practice;
- Define requirements - to outline the solution of the defined problem or rather outline an artefact and draw out requirements of the artefact;
- Design and develop artefact - to create an artefact that addresses the defined problem and satisfy the defined requirements;
- Demonstrate artefact - to prove the feasibility of a developed artefact through its application to a comprehensive case;
- Evaluate artefact - to determine the capability of an artefact to solve the problem and the extent to which it meets the requirements.

The research is further complemented with the help of the methodology for constructing composite indicators which can be seen as an inner cycle. The OECD et al. (2008) provided detailed descriptions of methods that can be used for composite indicator construction together with guidelines and the checklist for the development process of a composite indicator. The process of constructing a composite indicator consists of ten steps: theoretical framework, data selection, imputation of missing data, multivariate analysis, normalization, weighting and aggregation, robustness and sensitivity, back to the data, links to other indicators, and visualization of the results.

Each of the methodologies consists of a series of phases that can help researchers to design and carry out research of high quality. Whether methodology for constructing composite indicators corresponds to the method framework for design science research is checked in such a way that the steps and procedures of methodology for constructing composite indicators are mapped to what the method framework for design science research requires under the design science paradigm. Therefore, the activities of the method framework for design science research are mapped with the steps of the methodology for constructing composite indicators, considering the method framework for design science research as the outer cycle and the methodology for constructing composite indicators as the inner cycle of conducting the research (see Table 1.1). Furthermore, it seems that the developed solution of the problem is firstly demonstrated and then evaluated according to the method framework for design science research, while it is firstly evaluated and then demonstrated according to the methodology for constructing composite indicators. Demonstration of an artefact in DSR should have to include an application of the artefact to solve a practical problem, and the artefact evaluation should have to include an assessment of how well an artefact solves a problem (Peffers et al., 2007; Johannesson & Perjons, 2014). Since DSR represents an external cycle in the research, the application of a composite indicator in a specific situation, to check whether the developed composite indicator solves the problem

Table 1.1: The mapping of the activities of the Method framework for design science research and steps of the Methodology for constructing composite indicators.

	Activities of the Design Science Research				
	Explicate problem	Define requirements	Design and develop artefact	Demonstrate artefact	Evaluate artefact
Steps of the Methodology for constructing composite indicators	-	1. Theoretical framework 2. Data selection	3. Imputation of missing data 4. Multivariate analysis 5. Normalisation 6. Weighting and aggregation	10. Visualisation of the results	7. Robustness and sensitivity 8. Back to the data 9. Links to other indicators

at all, is more appropriate for the DSR's demonstration phase, while the robustness analysis and sensitivity analysis are more appropriate for the DSR's evaluation phase.

As a result of successfully mapping DSR activities with the steps of the methodology for constructing composite indicators, scientific contribution to the development of composite indicators is ensured and practical application of DSR in information systems research is realized.

The consolidation of the mentioned methodologies provided the development of the new artefact, or rather, a composite indicator along with the knowledge about it. The artefact can be defined as an object created by a human for the purpose of solving a problem in practice – in short, a solution to a practical problem (Johannesson & Perjons, 2014). Also, an artefact can be differently classified. Although the artefact classification defined by March and Smith (1995) is the most prevalent one within academia, Johannesson and Perjons (2014) stated that the classification of artefacts proposed by Offermann et al. (2010) truly reflects what practitioners in information technology (IT) and information systems would recognize as key artefact types. This pragmatic classification consists of eight artefact types (i.e., system design, method, language/notation, algorithm, guidelines, requirements, pattern, and metric) and it can be considered as more practical as well as more easily applicable than abstract classification posed by March and Smith (1995) which classifies artefacts into constructs, models, methods, and instantiations. Given that composite indicators are similar to mathematical or computational models, the artefact of the research can be classified as the 'metric' type according to a pragmatic artefact classification.

Also, it is important to highlight how to ensure a scientific contribution according to the design science research method framework. Unlike design, which produces working solutions to problems that are relevant only to a local practice, design science produces and communicates new knowledge that is of general interest. Therefore, Johannesson and Perjons (2014) identify three additional requirements for design science that arise from the different purposes of design and design science, and these are:

1. To create new knowledge relevant to a global practice, it is necessary that rigorous research methods are used within this research;

2. To ensure that the results produced are well-established and original, they must be related to an existing knowledge base;

3. The results obtained should be shared with practitioners and researchers.

Apart from mentioned methodological theories, other relevant sources used among others for defining concepts related to the subject, purpose, and theoretical framework are: previous studies related to the subject quality of open (meta)data, ISO 14721 Space data and information transfer systems — Open archival information system — Reference model (ISO 14721) by the International Organization for Standardization (ISO), ISO/IEC 11179 Information technology — Metadata registries (ISO/IEC 11179) by ISO and the International Electrotechnical Commission (IEC), ISO 9000:2015 Quality management systems — Fundamentals and vocabulary

(ISO 9000:2015) by ISO, ISO 19115-1:2014 Geographic information — Metadata (ISO 19115-1:2014) by ISO, Data Catalog Vocabulary (DCAT) by the World Wide Web Consortium (W3C), documentation for the various data management systems, different reports, etc (Neumaier et al., 2016; Kubler et al., 2018; Kučera et al., 2013; Reiche & Höfig, 2013; ISO, 2012; Consultative Committee for Space Data Systems, 2019; ISO and IEC, 2019; ISO, 2014, 2015; W3C, 2020; Lisowska, 2016; Simperl et al., 2014).

In accordance with the above mentioned, the conceptual framework of the research is proposed (see Figure 1.1). The conceptual framework can be considered as a logical structure that represents the visualization of how ideas are related to each other within the research through activities of the method framework for design science research. Since it is a logical structure, the activities are logically related through input-output relationships. It should be clear that activities are not temporally ordered. Hence, research is always conducted in an iterative way. The conceptual framework of the research is constructed by three main visual elements: ellipse shape represents the sources of data and information; rectangle shape with vertical lines illustrates the processes, and rectangle shape depicts the results of the processes. There is also an initial version of this framework that is refined later (Šlibar, 2019).

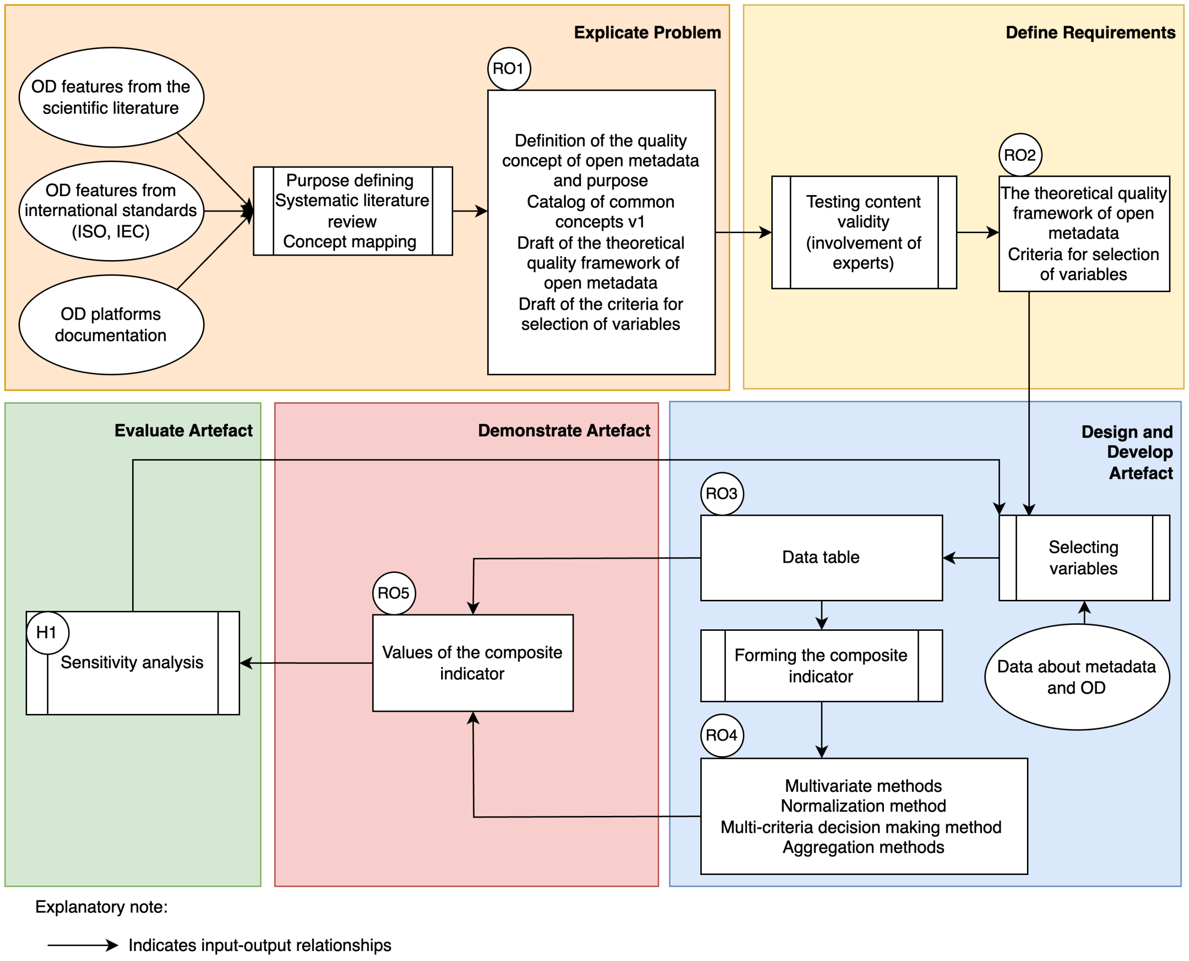

Explicate Problem
OD features from the scientific literature
OD features from international standards (ISO, IEC)
OD platforms documentation
Purpose defining
Systematic literature review
Concept mapping
RO1
Definition of the quality concept of open metadata and purpose
Catalog of common concepts v1
Draft of the theoretical quality framework of open metadata
Draft of the criteria for selection of variables
Define Requirements
Testing content validity (involvement of experts)
RO2
The theoretical quality framework of open metadata
Criteria for selection of variables
Evaluate Artefact
H1
Sensitivity analysis
Demonstrate Artefact
RO5
Values of the composite indicator
Design and Develop Artefact
RO3
Data table
Selecting variables
Data about metadata and OD
Forming the composite indicator
RO4
Multivariate methods
Normalization method
Multi-criteria decision making method
Aggregation methods
Explanatory note:
Indicates input-output relationships

Figure 1.1: Conceptual framework of the proposed research (Legend: RO - Research Objectives, H - Hypothesis, OD - Open Data, ISO - International Organization for Standardization, IEC - International Electrotechnical Commission).

CHAPTER 2

OPEN DATA

To address the recognised problem of metadata quality in open datasets, it is crucial to understand it. Therefore, it is important to first define all relevant concepts related to open data and find a solid basis for them in existing standards. Also, the scope of this research is defined in this chapter. As data portals are an essential part of open government data initiatives, this chapter provides an overview of the software, in particular the data management systems, on which these portals are based (see subchapter 2.1). Besides, metadata are described in more detail, including a brief historical development (subchapter 2.2.1), their basic components (subchapter 2.2.2), and existing standards (subchapter 2.2.3). The role and importance of metadata quality is also emphasised (see subchapter 2.2.4).

The analysis of existing definitions, conducted by the author, has shown that there is no generally accepted definition of open data, but the characteristics of such data were identified, as well as their importance considering their appearances in existing definitions (Šlibar et al., 2018). There are only a few definitions in the academic literature, and they are derived from the definitions of nonprofit organisations (e.g., Open Knowledge International), institutes (e.g., McKinsey Global Institute), or governments (Attard et al., 2015; Pasquetto et al., 2017). All content formats in digital form (text, image, audio, video, etc.) are considered as data in the context of open data. Open data are data that everyone is free to use, reuse, and redistribute, and that are subject to the highest requirements of preservation of provenance and openness (Šlibar et al., 2018). Moreover, open data should be publicly available through a public server and published in a machine-readable, preferably non-proprietary format (Šlibar et al., 2018; Attard et al., 2015; Pasquetto et al., 2017). A prerequisite for the existence of open data is the existence of an information infrastructure. Open data are available via data portals or so-called open data portals, whose main purpose is to provide a data catalogue. Hence, it could be said that these portals represent interfaces for accessing open data repositories that facilitate the process of data collection, publication, and distribution that should take place in one standardised way.

As discussed in the previous chapter 1, the relatively recent availability of open datasets (where open data is understood according to the definitions established by Open Knowledge International[1], the Open Data Institute[2], and the Open Data Policy Lab[3]) has left the harmonisation of terminology in the field of open data incomplete, resulting in a blending of fundamental concepts. In addition to defining basic concepts (see Glossary), it is also necessary to build upon existing standards, norms and other relevant documents to establish a solid foundation for the field of open data. The Reference Model for an Open Archive Information System (OAIS)

[1] `https://opendefinition.org/`

[2] `https://theodi.org/news-and-events/blog/what-is-open-data/`

[3] `https://opendatapolicylab.org/faq/`

has proven to be relevant and suitable for preservation (Consultative Committee for Space Data Systems, 2019; Lavoie, 2014; Zeng & Qin, 2022). In particular, it is one of the outcomes of the National Aeronautics and Space Administration's (NASA) Consultative Committee for Space Data Systems, whose work aimed to develop formal standards for the long-term preservation of digital data from space missions (Lavoie, 2014). The OAIS Reference Model was adopted in 2003 as ISO's international standard ISO 14721:2003 Space data and information transfer systems — Open archival information system — Reference model (ISO 14721:2003) and following its revision, a new updated version was published as ISO 14721:2012 Space data and information transfer systems — Open archival information system — Reference model (ISO 14721:2012) (ISO, 2012). The primary objective of the OAIS Reference Model is to provide a comprehensive and coherent framework for the description and analysis of digital preservation issues. It is intended to provide a solid foundation for future standardisation efforts and to serve as a reference point for those who seek to develop digital preservation products and services (Lavoie, 2014). As stated by Consultative Committee for Space Data Systems (2019), this reference model is suitable for all disciplines and organisations involved in the preservation and provision of digital information, whether they are already doing so or planning for the future. It is important to note that this reference model may not match all known terms of a particular discipline (e.g., traditional archives, digital libraries). It is already expected that each discipline or organisation will need to align some of its commonly used concepts with those outlined in the OAIS Reference Model (Consultative Committee for Space Data Systems, 2019). Since open data metadata is the focus of this study, only those concepts from the field of open data that are relevant in the context of this research are mapped alongside those listed in the OAIS Reference Model (see Table 2.1).

The scope of this research is limited to open public sector data, as the beginnings of the open data movement are linked to governments' awareness of the importance of making public sector information available to the public. Open public sector data, referred to as open government data, is defined as data collected by the public sector within its jurisdiction that can be FREELY reused for ANY purpose (Varga & Vračić, 2015; Musa et al., 2015). In addition to affecting innovation potential, the open data of the public sector can increase public sector transparency and openness, citizen involvement, and enable better law enforcement (Varga & Vračić, 2015; Musa et al., 2015).

Furthermore, Zuiderwijk et al. (2012) have stated that open data should not only be seen as a static product, but as an ongoing process. This perspective recognises that new applications and uses of open data can lead to new insights that may generate new approaches to using open data. The process of realising the potential of open data consists of the following five main steps (Zuiderwijk et al., 2012):

1. Creating data;

2. Opening data;

Table 2.1: The mapping of the Open Archival Information System (OAIS) Reference Model concepts and Open Data (OD) concepts.

OAIS Reference Model concept	OD concept
Representation Information	Semantic meaning
Content Information	Content of resource / Content of distribution
Content Data Object	Resource / Distribution + Metadata
Metadata	Metadata
Data	Data
Information	Subset of data
Information Object	Data + Semantic meaning
Data Management Functional Entity	Data management system (DMS), e.g., CKAN, Socrata
Open Archival Information System (OAIS)	OD portal
Archival Information Package (AIP)	Dataset / Catalogue record
Archival Information Collection (AIC)	All available datasets / Catalogue records
Archive	Publisher
Access Software	DMS's application program interface (API), e.g., CKAN API, SODA API
Access Functional Entity	Feature of OD portal for accessing data
Access Rights Information	Collection of metadata which describes access rights, e.g., licences
Ad-hoc Order	Hypertext Transfer Protocol (HTTP) GET request
Consumer	OD user
Data Dictionary	Controlled vocabulary
Descriptive Information	Metadata that make finding specific dataset easier, e.g., keywords, categories, publishers
Dissemination Information Package (DIP)	Response of HTTP GET request
Producer	Creator
Provenance Information	Provenance metadata
Reference Information	Metadata for identifying datasets, e.g., id, uniform resource identifier (URI)

3. Finding open data;

4. Using open data;

5. Discussing and providing feedback on open data.

In order to make the data easily findable and usable, which are essentially the only fully user-oriented steps to high-level representation in the open data process, quality information must be available (e.g., title, description, licence). This data-related information is called metadata. Clearly, a well-crafted description of the data increases the chances of discovering and using them to unlock their full potential. Consequently, the quality of metadata plays a pivotal role in realising the expected benefits of open government data (Dawes & Helbig, 2010; Zuiderwijk et al., 2012; Šlibar & Mu, 2022).

2.1 Data management systems for open government data

This subchapter begins by describing the concept of data management and emphasising its role and importance for organisations. The focus of this subchapter is on the systems underlying open data portals and their features, with a particular emphasis on the feature that plays a key role in data extraction from the end user's perspective.

Data management encompasses the development, execution and monitoring of strategies, policies, initiatives and procedures aimed at delivering, regulating, protecting and enhancing the value of data and information throughout their lifecycle (DAMA International, 2017). It encompasses a broad spectrum of activities ranging from informed decision-making to extracting strategic value from data to technical implementation. Therefore, data management requires both technical and non-technical, or more precisely, business knowledge and skills (DAMA International, 2017; Watson, 2016). Effective data management enables organisations to understand and satisfy their own information needs as well as those of other stakeholders. It encompasses the acquisition, storage, protection and preservation of data and ensures the quality, protection and confidentiality of data and information. It also protects against unauthorised or inappropriate access, manipulation or use of data and information and optimises data usage to increase the value of the data (DAMA International, 2017; Watson, 2016). As already mentioned, it is important to emphasise that data management goes beyond mere data administration. It encompasses the comprehensive management of data throughout its lifecycle to maximise its potential value.

A data management system is a software or set of tools designed to facilitate efficient data management (Watson, 2016). It serves as the backbone for storing, organising and manipulating data and thus forms the basis for managing data. A DMS enables data governance, data integration, data security and data quality control. Essentially, it automates many data management tasks, streamlines processes and makes data more accessible to authorised users. A notable feature of DMS is its role in metadata management, which involves maintaining comprehensive information about resources and making it searchable, discoverable, understandable, usable, and reusable (Ali et al., 2022; Braunschweig et al., 2012). The term open data management system is often used interchangeably with a term such as open data platform, especially in the context of open data initiatives, as can be seen in many existing studies (Šlibar & Mu, 2022; Neumaier et al., 2016; Milić et al., 2018).

Open data portals are powered by data management systems that enable data cataloguing, metadata management, data visualisation, an application programming interface (API), user support and many other features (Ali et al., 2022; Braunschweig et al., 2012). These systems make a decisive contribution to promoting the transparency and openness of government data. In order to identify data management systems that are frequently mentioned and used in the implementation of open data portals, a review of the relevant academic and professional literature on this topic is conducted. CKAN[4] , DKAN[5], JKAN[6], Junar[7], Magda[8], OpenDataSoft[9], and Open Data Platform powered by Socrata[10] (since its acquisition by Tyler Technologies in 2018, Socrata now operates as Tyler's Data and Insights division, continuing to leverage the software it was built upon) have established themselves as widely used data management systems for open data portals (Milić et al., 2018; Lisowska, 2016; Simperl et al., 2014; Bogdanović, Veljković, Frtunić Gligorijević, Puflović, & Stoimenov, 2021; Neumaier et al., 2016; *Top 16 Open source Data Portal Solutions for Open Data Publishing*, 2019). These systems offer various features, as already described, and each of them has its strengths and weaknesses. It is therefore up to organisations to choose the system they consider best suited to their needs. In addition to identifying widely used data management systems, examples of portals utilising these systems are also sought (see Table 2.2).

Table 2.2: Open data (OD) portals powered by widely used data management systems (DMSs).

DMS	DMS-based OD portal example
CKAN	https://www.data.gov/developers/apis https://open.canada.ca/en/access-our-application-programming-interface-api https://data.gov.hr/moodle/course/view.php?id=11
DKAN	https://data.cambridgeshireinsight.org.uk/about https://data.gov.jm/story/using-jamaicas-open-data-portal
JKAN	https://avaandmed.eesti.ee/instructions/api-manuall https://data.sandiego.gov/open-source/
Junar	https://data.cityofpaloalto.org/developers/
Magda	https://data.gov.au/api/v0/apidocs/index.html
OpenDataSoft	https://opendata.vancouver.ca/api/explore/v2.1/console https://linc.osbm.nc.gov/api/explore/v2.1/console
Open Data Platform powered by Socrata	https://dev.socrata.com/foundry/data.cityofnewyork.us/vx8i-nprf https://dev.socrata.com/foundry/data.cityofchicago.org/9xs2-f89t

The most common approach to enabling user access to open data is through APIs (González-Mora et al., 2023; Daga, Panziera, & Pedrinaci, 2015). They are an important feature of data management systems as they provide convenient access to data for interested stakeholders,

[4] https://ckan.org/

[5] https://dkan.readthedocs.io/en/latest/introduction/overview.html

[6] https://jkan.io/

[7] https://www.junar.com/

[8] https://magda.io/

[9] https://www.opendatasoft.com/en/

[10] https://www.tylertech.com/products/data-insights/open-data-platform

including users of open data. More precisely, the API serves as a significant intermediary that enables seamless communication and data exchange between websites and software or application (Jacobson, Brail, & Woods, 2012; Yang, Wittern, Ying, Dolby, & Tan, 2018). By sending an API request, the client system provides a uniform resource locator (URL) and a hypertext transfer protocol (HTTP) method, including header, body and parameters (depending on the method). Endpoints, specific URLs, work with HTTP methods to communicate with third-party services and users. API scraping, the process of extracting data, provide a structured and reliable method for accessing web applications, databases and online services (González-Mora et al., 2023; Khder, 2021). Unlike traditional web scraping, API calls interact directly with the backend of a service, ensuring targeted data retrieval (Khder, 2021). APIs allow direct access to specific data subsets via specific endpoints, so there is no need to navigate through extensive raw code or hypertext markup language (HTML) structures. The data collection process via an API involves initiating a request, authenticating via methods such as API keys, acquiring structured information (e.g., JSON or XML) and manipulating the data to fulfil programmatic requirements (Yang et al., 2018). This approach enables efficient data retrieval, saves bandwidth and ensures access to the most up-to-date data (Dietrich et al., n.d.).

All data management systems recognised as widely used for open data portals offer the possibility to download metadata via APIs that are necessary for using and analysing open data (González-Mora et al., 2023; Máchová & Lněnička, 2019). In addition, the metadata available via APIs depends on the underlying metadata models and thus on the schemas of the data management systems. As already addressed within subchapter 1.1 (page 4), it should be possible to assess the quality of metadata for open datasets independently of the underlying data management system of the portal, which has a direct impact on the generalisation of research results. The current inconsistency and diversity of metadata schemas in different data management systems also affect the automation potential of various processes, such as harvesting data from one portal to another or assessing metadata quality across different portals.

Inspired by previous studies that addressed the problem of different underlying metadata schemas of data management systems and emphasised the importance of harmonising metadata fields of such systems (Neumaier et al., 2016; Milić et al., 2018; Assaf, Troncy, & Senart, 2015), the author has attempted to map together the metadata fields of different data management systems that are relevant for open data initiatives. The source for the available metadata fields and their meaning was the official DMS API documentation and the actual metadata retrieved from open data portals (based on different DMSs) through APIs. The mapping of the metadata fields of the different data management systems was based on their semantic meaning.

During the mapping process, shortcomings of this approach are identified, which are described below. Firstly, data management systems are designed so that those who wish to use them to implement open data portals have the freedom to change the properties of existing metadata fields (whether they are required/mandatory or not) as well as to include other custom fields. Similarly, portals can be implemented in such a way that anyone who has the right to publish data on them can provide additional metadata fields. As a result, the same metadata field may have different properties (another term often used in practise for ‘property’ is ‘key’) on different portals based on the same data management system. It is to be expected that both the number and the structure of the fields in the metadata models will continue to vary.

Due to the above-mentioned shortcomings in the mapping of metadata fields of different data management systems and the conclusions of some other researchers, it was decided to harmonise metadata fields according to the relevant metadata standards (as described in subchapter 2.2.3). According to Milić et al. (2018), the solution to discrepancies in the metadata model could involve the use of a uniform metadata standard and the creation of appropriate mappings between metadata fields. Since the performed mapping of metadata fields from data management systems relevant for open data initiatives is not used as a basis for constructing the theoretical framework and developing a composite indicator, it is not further elaborated upon in this thesis.

2.2 Metadata

In this subchapter, a brief overview of the development of metadata is given, followed by definitions of the basic metadata components. In addition, an overview of metadata standards is given based on their intended use. A harmonisation of the metadata fields of two relevant metadata standards that define data structures and semantics while providing the ability to capture information about data sets is also provided. This serves as a basis for developing a theoretical framework and consequently a composite indicator. The subchapter concludes with an introduction to assessing the quality of metadata in open datasets.

2.2.1 A concise history of development

Metadata, or data describing information-bearing entities, can take various forms in our daily lives. The way we organise and describe information-bearing entities has remained consistent over time, despite changing methods and technologies. From handwritten catalogues to modern web services, this task has become essential due to our complex world and information overload. Libraries, archives, and museums have traditionally governed their resource organisation using rules and standards, with library cataloguing practises dating back to the 19th century (Zeng & Qin, 2022). These practises included the creation of cataloguing rules by Antonio Panizzi, Charles Coffin Jewett’s library building efforts, and Charles Ammi Cutter’s rules for a printed dictionary catalogue as described by Zeng and Qin (2022). These practises led to rich bibliographical descriptions using classification schemes and lists of subject headings. The emergence of the Anglo-American Cataloguing Rules (AACR) (Francis, Haider, Foskett, & Estabrook, n.d.) and the MAchine-Readable Cataloging (MARC) format (Furrie, 2003) played an important role, but the rise of the Internet led to new challenges in the development of metadata.

In the pre-Internet era, the focus was mainly on physical information objects such as books and journals. Cataloguing these objects was labour intensive and technology could not easily replace human cataloguers. However, with advances in information technology, libraries moved to online systems, enabling various cataloguing services and the creation of online union catalogues and shared systems.

In the Internet age, the importance of library catalogues to society remains, but new information technologies bring with them changing requirements. The emergence of distributed repositories on the Internet triggered the development of new guidelines and architectures. The development of metadata standards gained momentum in the 1990s as various communities took up the challenge of managing digital resources (Pomerantz, 2015; Zeng & Qin, 2022).

Over the years, the scope of metadata standards grew, including for digital collections, and they often built on earlier standards. Large metadata vocabularies emerged, such as Schema.org, which aimed to make web content more accessible to search engines.

Metadata research and practice have expanded and evolved in remarkable ways in the last 25 years. National and international digital libraries (e.g., Europeana[11]) have been established. Concepts such as the Semantic Web, Open Data, Big Data and Linked Open Data have revolutionised metadata research and applications (Riley, 2017; Zeng & Qin, 2022). Metadata standards have adapted to these changes and the growing importance of data in today's world. The ever-increasing amount of data in our daily lives has led to the fight against misinformation, with metadata research and implementation playing a crucial role in it.

2.2.2 Basic metadata components

As Zeng and Qin (2022) have explained with a simple example involving labels on food packaging and beverage containers that contain information about ingredients and nutritional content, we can think of these labels as a form of metadata. Metadata descriptions employ paired property-value statements to precisely outline the attributes of a particular item, as is the case with these food labels. A property, sometimes called an attribute or element, gives meaning to a data value. For example, the property 'energy kcal' allows the consumer to grasp the meaning of the data value '190' on this food label. Maintaining a consistent structure and format in food labelling is of paramount importance.

Similarly, the concept of metadata can be illustrated in the context of open datasets. It should be noted that each domain has unique properties that require structured data specifically designed for describing them (Zeng & Qin, 2022; Pomerantz, 2015). Metadata descriptions for open datasets can provide information about the publisher of the dataset, update frequency, available data formats, terms of use, data sources and more. In this way, users can select the datasets they need with confidence in their authority and quality.

To distinguish and fully understand the basic components of metadata, we present the following key aspects based on the Dublin Core Metadata Initiative (DCMI) Abstract Model (Zeng & Qin, 2022; Powell et al., 2007):

[11] `https://www.europeana.eu/en`

- A description set, also called a metadata record, comprises one or more metadata descriptions, each of which serves to describe a single entity.
- A metadata description comprises one or more statements about a single entity and may include a uniform resource identifier (URI) for the described entity.
- A metadata statement is an essential element of the metadata that represents a property together with its corresponding value. The alternative term used in this research for metadata statement is metadata field.

Whether a metadata record describes the entire collection or units within the collection, the fundamental descriptions upon which a metadata repository is built vary (Zeng & Qin, 2022). Accordingly, Zeng and Qin (2022) described four levels of granularity in relation to the fundamental descriptions: ‘item-level’, ‘collection-level’, ‘dataset-level’, and ‘resource decomposition’. The term ‘dataset-level’ metadata is used when the dataset represents the fundamental unit of description which is the case in most of the metadata standards for scientific data. This also applies to open data initiatives. A dataset class is defined as a class within an ontology in the W3C’s Data Catalogue Vocabulary - Version 2 (DCAT 2), which also includes the following definition: ‘a collection of data published or curated by a single agent, and available for access or download in one or more representations’ (W3C, 2020). Eventually, the object of observation in this research is an open dataset, or more precisely, metadata for a dataset.

2.2.3 Metadata standards

Usually, the term metadata standard is used to refer to a specific metadata element set and/or schema that has been officially recognised by a national or international standards organisation, a specific community, or a professional association (Zeng & Qin, 2022). These standards are widely recognized and play a crucial role in establishing consistency and effectiveness in managing metadata. Therefore, Zeng and Qin (2022) have divided metadata standards into four categories based on their intended use:

1. Data structure standards: These standards, often referred to as element sets or metadata vocabularies, define the structures and semantics of data, e.g., Dublin Core standards, Schema.org, or Data Catalog Vocabulary.
2. Data content standards: They serve as a guide for metadata creation and cataloguing, e.g., Resource Description and Access.
3. Data value standards: These include knowledge organisation systems (KOSs) commonly known as value vocabularies and sometimes value encoding schemes within a metadata vocabulary specification, e.g., ISO 639 Codes for individual languages and language groups (ISO 639), DCMI Type Vocabulary.

4. Data exchange standards: In the context of data exchange and communication, these standards are often referred to as different formats, e.g., Extensible Markup Language, Resource Description Framework.

An initial harmonisation of the metadata fields of data management systems relevant to open data initiatives (as described in subchapter 2.1) revealed that the discrepancy in underlaying metadata schemas of data management systems and thus of open data portals could be handled by using one or more common metadata standards and an appropriate mapping between metadata fields.

Harmonisation therefore takes place at the level of metadata standards, together with a clear description of the meaning of each metadata field, i.e., a metadata statement. In this way, anyone wishing to use metadata fields for any purpose, including quality assessment, can map the fields from the portals to those defined according to the standard(s). The standards that have proven suitable for this research are those that define data structures and semantics and provide the ability to record information about datasets, specifically Data Catalog Vocabulary - Version 2 (W3C, 2020) and ISO/IEC 11179-7:2019 Information technology — Metadata registries — Part 7: Metamodel for data set registration (ISO/IEC 11179-7:2019) (ISO and IEC, 2019). When mapping the standards metadata fields, some other parts of ISO/IEC 11179 were also used for the sake of completeness, in particular ISO/IEC 11179-3:2013 Information technology — Metadata registries — Part 3: Registry metamodel and basic attributes (ISO/IEC 11179-3:2013) and ISO/IEC 11179-6:2015 Information technology — Metadata registries — Part 6: Registration (ISO/IEC 11179-6:2015) (ISO and IEC, 2013, 2015). The results of this mapping are stored on the Harvard Dataverse research data repository as a dataset (Šlibar, 2024b).

It is important to mention that the mapping of metadata standards fields was made at the end of 2022. Since then, a new version of DCAT has been published and all parts of ISO/IEC 11179 have been updated. As described in (W3C, 2023), Data Catalog Vocabulary - Version 3 (DCAT 3) builds on DCAT 2 without making it obsolete. DCAT 3 maintains the DCAT namespace and ensures backward compatibility with DCAT 2 by retaining the existing terms. Furthermore, DCAT 3 provides more flexibility and introduces new classes and properties while preserving the definitions of the existing terms. While the adoption of new implementations is in DCAT 3, existing implementations are not required to upgrade unless they intend to utilise the new features. It is important to note that this does not apply to ISO/IEC standards. A comparison was made between the adopted part of ISO/IEC 11179-7:2019 and the new version called ISO/IEC 11179-33:2023 Information technology — Metadata registries — Part 33: Metamodel for data set registration (ISO/IEC 11179-33:2023) (ISO and IEC, 2023). However, there was no need to update the previously conducted field mapping based on the newly emerged changes.

2.2.4 Metadata quality

The quality of metadata must be defined in order to be able to measure and improve it. As already addressed in subchapter 1.1 (page 4), none of the attempts to clarify the definition of quality, data quality or metadata quality have been widely accepted to date. As Juran and DeFeo

(2010) noted, capturing the necessary depth of meaning in a short sentence is a challenge when it comes to quality. In search of a concise definition that can be universally applied in different scenarios, Juran and DeFeo (2010) revised the earlier definition of quality, previously referred to as ‘fitness for use’, into a new definition ‘fitness for purpose’. This means that every product or service must have the right features to satisfy customer needs and must be delivered with few defects to be effective, efficient and achieve superior business performance. Similarly, ISO (2015) defines quality as ‘degree to which a set of inherent characteristics of an object fulfils requirements’.

However, there is a consensus that the quality of metadata, like quality, is a multidimensional concept. It is therefore important to include all relevant aspects in order to determine whether and to what extent something is of quality. The quality can be assessed at different levels, namely at the collection, metadata record and metadata statement, using different indicators to capture all relevant aspects (i.e., dimensions) (Zeng & Qin, 2022). Quality issues can occur at different stages of the metadata lifecycle, from schema design to the creation of metadata descriptions, as well as during data conversion and/or data aggregation. Thus, the evaluation of metadata can be focused on the assessment of processes associated with metadata or on the metadata itself (as is the case in this research).

According to Zeng and Qin (2022), the quality of metadata is usually measured at the metadata record level by examining an entire population or a sample from an existing repository. Also, the quality of the metadata from open data portals was assessed at the metadata record level in this research. When data is submitted to a repository, the associated metadata can be checked manually or automatically. This process of quality control has long been common practise in the library communities. Although this process is often impractical in many digital repository development projects, due to budget and resource constraints (Zeng & Qin, 2022).

In order to identify all relevant aspects that need to be considered when measuring the quality of metadata in open datasets, scientific and professional literature, as well as expert judgement, were taken into account. The results are presented in subchapter 4.1 (page 50).

CHAPTER 3

METHODOLOGY

The previous chapters provide a base for designing and developing a theoretical framework as well as a composite indicator within the scope of this research. As is already stated in the subchapter 1.3 (page 8), this research is carried out through several activities and related steps involving predominantly quantitative and, to a minor extent, qualitative research methods, following the method framework for design science research and the methodology for constructing composite indicators (OECD et al., 2008; Johannesson & Perjons, 2014).

This chapter is divided into two subchapters. Subchapter 3.1 contains criteria for selecting data that are used to develop the composite indicator, whereas subchapter 3.2 follows the five activities of the method framework for design science research and provide detailed descriptions of research methods applied within each activity.

3.1 Data description

The prerequisite for data collection is the selection of the portal from which the data will be collected. As stated in the subchapter 1.1 (page 4), the quality assessment of open metadata should not be determined by a data management system. Therefore, it is necessary to select at least two portals that are built upon different data management systems. The portals should also offer a large quantity and variety of open datasets. Data are collected from the official portal for European data[1], powered by CKAN, and from the central source of Australian open government data[2], powered by Magda. These portals were chosen not only because they are based on different data management systems, but also because they host thousands of open datasets published by international, federal, national, regional, and local authorities, among others. In addition, the metadata for all open datasets on these portals is available in English, which is not always the case for national, local or other specific open data portals.

Before collecting the data, it is also necessary to determine the sample size needed for statistically significant results for the population since the selected portals contain thousands of datasets. Confidence intervals can be used to estimate a few population parameters, one of which is population proportion. To determine the sample size, a formula for a confidence interval for population proportion is used (Holmes, Illowsky, Dean, & Hadley, 2017). The smallest sample size required to obtain a margin of error of at most ±2% at a 95% confidence level is determined, which is 2,401 datasets from each selected OGD portal.

In addition, a separate R script is written for each portal to collect the data. For randomly selected lists of datasets (i.e., JSON array of 10 datasets), the corresponding API request is sent

[1] `https://data.europa.eu/`

[2] `https://data.gov.au/`

to fetch JSON data from a portal. Then, the retrieved lists are merged and stored in a .json file containing data about metadata from an open data portal. The functions from the jsonlite package are used for this purpose (Ooms, 2023). Since API requests are sent for randomly selected lists of datasets and each list contained 10 datasets, a total of 4,820 open datasets are retrieved. In addition, a check is made to see if there are duplicate datasets (those with the same identifier) in the sample. The identified duplicates (198 in total) are removed from the sample.

3.2 Research methods

Since the method framework for design science research under the paradigm of design science represents an outer cycle of conducting this research, this subchapter follows the five main activities of the method framework for design science research and provides detailed description of associated steps of the methodology for constructing composite indicators together with research methods used in each step.

3.2.1 Explicate problem

The conceptual framework of this research represents a logical structure in which activities are logically related through input-output relationships, as described in subchapter 1.3 (page 8). Basically, every activity requires input(s) and generates output(s). The first activity of the method framework for design science research is the explication of the problem. This activity needs the initial problem as input because it is about exploring the practical problem. Therefore, the problem should be clearly stated, generalisable, and its practical relevance should be justified (Johannesson & Perjons, 2014).

For the avoidance of doubt, it is important to define the basic terminology utilized in constructing a composite indicator in advance. The composite indicator has a theoretical, hierarchical structure behind it. The theoretical framework for the composite indicator is thus a hierarchical structure of concepts, at the top of which is the composite indicator, below which are elements arranged on one or more levels, and at the bottom of this structure are the individual indicators. Each element in the structure of the theoretical framework for the metadata quality of open datasets that can be computed as a partial composite indicator, is called a dimension. A dimension results from the aggregation of either individual indicators or other dimensions. Therefore, an element in this structure is considered a dimension as long as it is the result of an aggregation function. The element at the bottom of this structure, more specifically the individual indicator, represents the result of applying metrics to a property/attribute (i.e., a metadata statement). Metrics stand for the quantification of the quality properties of metadata.

A systematic literature review (SLR) is conducted, by the author, to provide a broad overview of the most recent advancements in open data evaluation (Šlibar et al., 2021). Two research questions from the paper relevant to this research are (Šlibar et al., 2021): ‘What (meta)data quality (sub)dimensions are used for OD assessment?’ and ‘How are (meta)data quality (sub)dimensions measured in studies that assess OD?'. The review procedure in this paper is designed based on how others (Attard et al., 2015; Kitchenham et al., 2010; Novak, Joy, & Kermek, 2019; Ruijer & Martinius, 2017) have applied the method and based on the guidelines proposed by Littell, Corcoran, and Pillai (2008); Petticrew and Roberts (2006). More details on framing the research questions, defining the exclusion criteria, determining the most relevant electronic sources, forming the search query, and selecting the most relevant papers according to identified research objectives can be found in the paper (Šlibar et al., 2021).

Apart from the SLR, for understanding and defining of the multidimensional phenomenon to be measured, determining the structure of the various sub-groups of the phenomenon, identifying the selection criteria for the underlying elements, and defining underlying elements, other scientific papers are used that were not included in the SLR (e.g., papers that refer to the quality assessment of data or metadata in general) and professional literature (e.g., international standards, various reports), as it is shown in subchapter 4.1.1 (page 50).

3.2.2 Define requirements

Define requirements activity extends the previous activity by outlining the artefact and elaborating its requirements (Johannesson & Perjons, 2014). This activity comprises two steps of methodology for constructing composite indicators: Theoretical framework and Data selection. The Theoretical framework step in the construction of a composite indicator is explained as a foundation that should be developed in a way that provides a basis for selecting and combining individual indicators into a meaningful composite indicator that satisfies the fitness-for-purpose principle, while the Data selection step involves checking the quality of the individual indicators, discussing their strengths and weaknesses, and creating a summary table of the data characteristics (OECD et al., 2008).

Since, according to Johannesson and Perjons (2014), requirements represent properties of an artefact that are considered desirable by stakeholders in a practice and guide the design and development of the artefact, opinions of different stakeholders in the open data ecosystem (Simperl et al., 2014) are taken into account in this activity. The criteria for selecting experts required them to belong to one of the following two groups: those within public administration, focusing on OD/IT strategists in the public sector; or those outside the public administration, focusing on open data users, primarily researchers. Additionally, the selection criteria stipulated that experts come from at least a few different countries. At first, it was planned to select only experts from the list of collaborators of two, now completed, international projects related to open data, which were funded by EU programs and in which the Faculty of Organization and Informatics Varaždin was involved: i) the project Twinning Open Data Operational (TODO)[3], funded by Horizon 2020 and involving eight partners from three EU Member States; ii) the thematic network Shared Standards for Open Data and Public Sector Information (Share-PSI 2.0)[4],

[3] https://cordis.europa.eu/project/id/857592

[4] https://cordis.europa.eu/project/id/621012

funded by the Information and Communication Technologies Policy Support Program (ICT PSP) and involving 40 institutions from different European countries. However, the initial sources of experts' contacts are extended by the author's participation in international scientific or professional conferences[5] (e.g., EU DataViz 2021), summer schools (e.g., Summer School of the TODO project) or some other events. Most of the experts who responded and agreed to participate have over 10 years of experience working with open data and have varied backgrounds. They have worked in, or have experience in, e-government, consulting, development, research, and the use of open government data. Currently, out of the total of 11 experts, six work in government organisations, while five are employed in the private sector or academia. Additionally, all experts come from different countries, including Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Romania, and Serbia, with the exception of two experts who are from Croatia.

In addition to the literature review in developing the theoretical framework and selecting the data, the Q methodology in combination with Lawshe's content validity ratio (CVR) is used (Brown, 1993; Lawshe, 1975).

The Q methodology is applied within this research for purpose of identifying, understanding, and categorizing experts' opinions as well as for grouping experts based on their opinions about statements, i.e., individual indicators. From the larger pile consisting of all the individual indicators proposed in the previous activity, a subset of individual indicators is selected, called the Q sample. This group of individual indicators is presented to the participants in the form of a Q sort. The opinions of experts on which individual indicators are relevant to a particular quality dimension of open data metadata are collected with the help of the EQ Web Sort tool, as explained below (Banasick, 2023b). The EQ Web Sort is software that allowed the author herself to set up an online Q sorting for a Q methodology. In addition, the EQ Web Configurator is used for setting-up and testing an online Q sorting task with EQ Web Sort (Banasick, 2023a). For each quality dimension of open data metadata, a separate online Q sorting[6] is designed. A different Q sample is prepared for each Q sort. The Q sample consists of individual indicators that are part of an observed dimension and random individual indicators that are part of other dimensions according to the literature. Besides, the Q sorting task involves pre-sorting (to make initial distinctions between the statements) and sorting steps (to make finer distinctions between the statements), as suggested by Lawshe (1975). This means that an expert can perceive each individual indicator as 'Essential', 'Useful but not essential', or 'Not necessary'. When pre-sorting, an expert should place the statements into three piles:

- A pile for statements that tended to be disagreed with labelled Not necessary;
- A pile for statements that tended to be agreed with labelled Essential;

[5] Conference participants could leave contact information by filling following form `https://forms.gle/jCANAwX4NeSTQiXj6`

[6] Link to qsorting tasks used for this research: `https://services.foi.hr/qsort/`

- A pile for the rest labelled Useful but not essential.

During the sorting step, an expert should order the statements onto spaces on a Q grid, a distribution close to a standard normal distribution. For each Q sorting, the Q grid is determined individually based on the number of statements and the maximum Q sort rank. So, statements that one considers as the most relevant (i.e., Essential) should be placed on the far-right side of the grid, while statements that one considers not to be relevant at all (i.e., Not necessary) should be placed on the far-left side. The design of the Q grids is determined using the R package qmethod (Zabala, Held, & Hermans, 2023).

The invitation to participate in testing the content validity of the proposed theoretical framework, i.e., to identify which individual indicators are relevant to a particular quality dimension of open data metadata, is sent to the experts by e-mail (see Appendix A). Documentation that is attached to the invitation email included: a document entitled ‘Definitions of Dimensions’, which contains a description of each dimension (as defined in subchapter 4.1.1, page 50); a document entitled ‘Descriptions of Metadata Fields’, which contains the mapping of metadata fields and their descriptions according to the metadata standards (as described in subchapter 2.2.3, page 23); a research paper by Brown (1993); and a project report by Hunnius et al. (2015). Expert responses collected through Q sorting tasks are seamlessly downloaded from the Firebase web host. Subsequently, manual data cleaning is performed, e.g., deleting redundant lines or characters, adding labels to the statements, etc. Further analysis of cleaned data is done in R using psychometric, xtable, psych, and qmethod packages (Fletcher, 2023; Dahl, Scott, Roosen, Magnusson, & Swinton, 2019; Revelle, 2023; Zabala et al., 2023).

Lawshe’s content validity ratio is calculated for the purpose of identifying relevant individual indicators, as explained below. It is considered that the experts have chosen an individual indicator as relevant to a dimension, if they rank it with a number greater than -1 on the Q grid. Assuming a binomial distribution, if there is an equal probability that the expert declares the statement to be relevant or not, it is to be expected that some of the experts will randomly agree that the statement is relevant (Lawshe, 1975; Ayre & Scally, 2014; Ermis-Demirtas, 2018). Critical CVR values are known from the literature(Lawshe, 1975; Ayre & Scally, 2014; Ermis-Demirtas, 2018). As suggested by Ayre and Scally (2014), exact binomial distribution is used to estimate probability that the observed or a larger number of experts could have assessed an individual indicator as a relevant one by chance. Also, the average value of the Q sort rank assigned to an individual indicator by the experts during the sorting step is determined. The one-sample t-test is used to test the hypothesis that the mean Q sort rank (i.e., average score) is significantly larger than 0. Consequently, an individual indicator is selected as relevant based on the following criteria:

1. If the resulting p-value of the exact binomial test or the p-value of the one-sample t-test is less than 0.1;

2. If an individual indicator is found to be relevant in more than one dimension according to

the previous criterion, it is assigned to only one dimension based on the mean Q sort rank, the one in which the mean Q sort rank is the highest and positive.

To perform a full analysis with the Q methodology, the number of factors / principal components (i.e., different profiles of responses, or rather, perceptions of the importance of combinations of statements) to be extracted should be determined in advance. The Horn's method, also known as parallel analysis (Horn, 1965), is used to determine the number of factors/principal components with the package psych in R (Horn, 1965; Field et al., 2012; Revelle, 2023). Complete Q methodology analysis is run in R using the qmethod package (Zabala et al., 2023).

3.2.3 Design and develop artefact

The third activity, Design and develop artefact of the design science research method framework, is about creating an artefact that addresses the defined problem and meets the defined requirements (Johannesson & Perjons, 2014). The artefact developed is a composite indicator based on the metadata quality framework for open datasets. The methodology for constructing composite indicators is applied following the guidelines of the Organisation for Economic Co-operation and Development and the Joint Research Center of the European Commission (OECD et al., 2008).

This activity begins with the process of selecting variables, which essentially involves three things. First, the metadata fields of each selected portal are mapped to the metadata fields included in the list created based on the metadata standards (Šlibar, 2024b). Not all metadata fields available through a portal API are matched against those specified in the metadata standards, but only those that are part of the individual indicators identified as relevant in the previous activity. Functions from the rjson package are used to load previously retrieved metadata from the portal (Couture-Beil, 2022).

Second, the scores of the relevant individual indicators are calculated for each retrieved dataset, as explained below. All individual indicators refer to a single metadata field, either at the dataset level or at the distribution level, except for the one that includes two metadata fields. To evaluate the existence of a metadata field at the dataset level, it is checked whether the value of a specific metadata field is filled in or not (as shown in Code snippet 3.1). The scores of the following individual indicators are calculated in the described manner: c11, c18, f20, f21, c_f22, f23, f24, c_f31, f33, u54, and u56[7].

The validity of the data format of the metadata field *Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date* (labelled as c12) is checked by comparing its value to the ISO 8601 Date and time — Representations for information interchange (ISO 8601) date format (ISO, 2019), as shown in Code snippet 3.2.

[7] The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are those that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

```
dataDoesntExists <- function(data)
{
        is.null(data) || is.na(data)
}
if(!(dataDoesntExists(row$dataset_last_change_date)))
{
        dataset_last_change_date_exists <- dataset_last_change_date_exists + 1
}
```

Code snippet 3.1: Evaluation of the existence of a metadata field at the dataset level.

```
isDate <- function(mydate, date.format = "%Y-%m-%d") {
        tryCatch(as.Date(mydate, date.format),
        error = function(err) {FALSE})
}
```

Code snippet 3.2: Evaluation of the validity of the metadata field *Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date* at the dataset level.

Since there is no defined format for the value of the field *Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier*, the conformity of the metadata field *Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier* (labelled as f34) is checked by attempting to access dataset in a portal using combination of a portal URL and the value of the field *Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier* (see Code snippet 3.3). Considering that access to a dataset may fail if a portal is down and unavailable, access to a dataset is attempted twice.

Conformity of the metadata field *Data_Set ->rights; Cataloged_Resource ->access_rights* (labelled as u55) is determined by comparing its value with the access rights or restrictions to distributions specified in the EU controlled vocabulary for access rights[8], as shown in Code snippet 3.4.

To check whether a license specified in the field *Data_Set ->rights; Cataloged_Resource ->license* is open (labelled as u58), a list[9] of licenses compatible with Open Definition is retrieved first. As this URL returns response in JSON format, the jsonlite package is used to retrieve the list and transform it into R list (Ooms, 2023). Then, it is checked whether the value set in the field *Data_Set ->rights; Cataloged_Resource ->license* matches any value from that list (see Code snippet 3.5).

[8] Link to the EU controlled vocabulary for access: `https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/access-right`

[9] Link to the list of licenses approved as conforming to the Open Definition: `https://licenses.opendefinition.org/licenses/groups/od.json`

```
tryCatch(
{
        GET(row$dataset_id, timeout(timeToWaitForResponseInSeconds))
        if(response$status_code == 200)
        {
                dataset_id_valid <- dataset_id_valid + 1
        }
}
,error = function(err)
{
        Sys.sleep(5)
        GET(row$dataset_id, timeout(timeToWaitForResponseInSeconds))
        if(response$status_code == 200)
        {
                dataset_id_valid <- dataset_id_valid + 1
        }
}
```

Code snippet 3.3: Evaluation of the conformity of the metadata field *Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier* at the dataset level.

```
accessRights <- read.csv("access_rights.csv")
numberOfMatches <- which(accessRights$codes == tolower(row$dataset_access_control))
if(length(numberOfMatches) > 0)
{
        dataset_access_control_conformal <- dataset_access_control_conformal + 1
}
```

Code snippet 3.4: Evaluation of the conformity of the metadata field *Data_Set ->rights; Cataloged_Resource ->access_rights* at the dataset level.

To evaluate the existence of a metadata field at the distribution level, first it is checked how many distributions attached to the dataset have the value of a specific filled in metadata field. Then, this number is divided by the total number of distributions attached to the dataset (see Code snippet 3.6). The scores of the following individual indicators are calculated in the described manner: c4, c13, i37, i42, r48, r51, and u61[10].

Conformity of the metadata field *Data_Set_Distribution ->format; Distribution ->format* (labelled as i39) is determined by checking whether the value of the field *Data_Set_Distribution ->format; Distribution ->format* of each distribution attached to the dataset is mentioned within the list of IANA media types[11] (see Code snippet 3.7). Afterwards, the number of dataset distri-

[10] The definitions of abbreviations for individual indicators are provided in Tables 4.2, 4.3, 4.4, 4.5, pages 56-59.

[11] Link to the most up-to-date and complete list of IANA media types: `https://www.iana.org/assignments/media-types/media-types.xhtml`

```
cleanedUpLicense <- function (license)
{
        license <- tolower(license)
        license <- str_replace_all(license, "\\.","")
        license <- str_replace_all(license, " ","")
        license <- str_replace_all(license, "-","")
        license <- str_replace_all(license, "_","")
}

url <- "https://licenses.opendefinition.org/licenses/groups/od.json"
openLicenses <- fromJSON(txt=url)

processedLicense <- cleanedUpLicense(row$dataset_license)

licenseList <- list()
for(index in 1:length(openLicenses))
{
        openLicense <- openLicenses[[index]]

        if(openLicense$id != "")
                licenseList <- append(licenseList, cleanedUpLicense(openLicense$id))
        if(openLicense$title != "")
                licenseList <- append(licenseList,
                cleanedUpLicense(openLicense$title))
        if(openLicense$url != "")
                licenseList <- append(licenseList, cleanedUpLicense(openLicense$url))
}

numberOfMatches <- 0

for(index in 1:length(licenseList))
{
        if(grepl(processedLicense, licenseList[index], fixed = TRUE))
        {
                numberOfMatches <- 1
        }
}

if(numberOfMatches > 0)
{
        dataset_license_open <- dataset_license_open + 1
}
```

Code snippet 3.5: Evaluation of the openness of the metadata field *Data_Set ->rights; Cataloged_Resource ->license* at the dataset level.

```r
numberOfResources <- function(resourceAttribute) {
        length(listFromAttribute(resourceAttribute))
}

dataframe_resource_releasedate_exists <- c(dataframe_resource_releasedate_exists,
numberOfResources(row$resource_releasedate)/row$number_of_resources)
```

Code snippet 3.6: Evaluation of the existence of a metadata field at the distribution level.

butions conforming to the official MIME types is divided by the total number of distributions of a dataset.

```r
ianaMediaTypes <- read.csv("iana.csv")
numberOfMatches <- 0
for(index in 1:length(ianaMediaTypes$Template))
{
        if(grepl(tolower(resourceFormat), ianaMediaTypes$Template[index]))
        {
                numberOfMatches <- 1
        }
}
if(numberOfMatches > 0)
{
        resource_format_iana <- resource_format_iana + 1
}
```

Code snippet 3.7: Evaluation of the conformity of the metadata field *Data_Set_Distribution ->format; Distribution ->format* at the distribution level.

The validity of the metadata field *Data_Set_Distribution ->access_url; Distribution ->access_URL* (labelled as r49) is determined by checking whether the value of the field *Data_Set_Distribution ->format; Distribution ->format* of each distribution attached to the dataset matches RegEx expression, as shown in Code snippet 3.8. The stringr package in R is used to check whether the field values match the RegEx expression (Wickham, 2023b). As with the previous individual indicators at the distribution level, the number of distributions in a dataset with a matching value is divided by the total number of distributions in a dataset.

There are three different ways (each being executed if one before fails) to verify whether the metadata field *Data_Set_Distribution ->format; Distribution ->format* is accurate (labelled as i38), as shown in Code snippet 3.9:

- First, it is checked whether the value of the metadata field *Data_Set_Distribution ->download_url; Distribution ->download_URL* ends up with the extension that matches the value of the field *Data_Set_Distribution ->format; Distribution ->format*.

```
isValidUrl <- function(string) {
        pattern <- "(https?|ftp)://[^ /$.?#].[^\\s]*"
        str_detect(string, pattern)
}

if(isValidUrl(accessUrl))
{
        resource_access_url_valid <- resource_access_url_valid + 1
}
```

Code snippet 3.8: Evaluation of the validity of the metadata field *Data_Set_Distribution ->access_url; Distribution ->access_URL* at the distribution level.

- Second, the GET request is sent to the URL written in the metadata field *Data_Set_Distribution ->download_url; Distribution ->download_URL* using httr package in R (Wickham, 2023a). The content type received in the response is compared to the value of the metadata field *Data_Set_Distribution ->format; Distribution ->format*.
- Third, if the content type contains value 'zip', the file from the metadata field *Data_Set_Distribution ->download_url; Distribution ->download_URL* is downloaded and unpacked using tools package in R (Hornik & Leisch, 2023). The extension of every unpacked file is checked against the value of metadata field *Data_Set_Distribution ->format; Distribution ->format*.

Afterwards, the number of distributions with matching formats attached to the observed dataset is divided by the total number of distributions of the observed dataset.

The openness (open / non-proprietary formats) of the metadata field *Data_Set_Distribution ->format; Distribution ->format* (labelled as i40) is determined by checking whether the value of the field *Data_Set_Distribution ->format; Distribution ->format* of each distribution attached to the dataset is mentioned in a predefined set of confirmed open/non-proprietary formats[12] (see Code snippet 3.10). As with other individual indicators at the distribution level, the number of distributions in a dataset with a matching value is divided by the total number of distributions in a dataset.

The openness (machine-readable file formats) of the metadata field *Data_Set_Distribution ->format; Distribution ->format* (labelled as i41) is determined by checking whether the value of the field *Data_Set_Distribution ->format; Distribution ->format* of each distribution attached to the dataset is mentioned in a predefined set of confirmed machine-readable file formats[13] (see Code snippet 3.11). As with other individual indicators at the distribution level, the number of

[12] The list of confirmed open / non-proprietary formats is created based on two sources: `https://github.com/opendatamonitor/odm.restapi/blob/master/odmapi/def_formatLists.py`, `https://gitlab.com/european-data-portal/edp-vocabularies/-/blob/master/edp-non-proprietary-format.rdf`

[13] The list of confirmed machine-readable file formats is created based on two sources:

```
if(grepl(tolower(paste(".",format,sep="")), tolower(url))
{
        resource_format_accurate <- resource_format_accurate + 1
}
else
{
        if(!dataDoesntExists(format) && !dataDoesntExists(resourceContentType))
        {
                if(grepl(tolower(format),tolower(resourceContentType)))
                {
                        resource_format_accurate <- resource_format_accurate + 1
                }
                else if(grepl("zip", resourceContentType, fixed = TRUE))
                {
                        dir = "downloadedResource.zip"
                        zippedFiles <- NULL
                        download.file(url, dir)
                        zippedFiles <- unzip(dir, list=TRUE)
                        for(fileNameIndex in 1:length(zippedFiles$Name))
                        {
                                if(tolower(file_ext(zippedFiles$Name[[fileNameIndex]])
                                == tolower(format))
                                {
                                        resource_format_accurate
                                        <- resource_format_accurate + 1
                                        break
                                }
                        }
                }
        }
}
```

Code snippet 3.9: Evaluation of the accuracy of the metadata field *Data_Set_Distribution ->format; Distribution ->format* at the distribution level.

distributions in a dataset with a matching value is divided by the total number of distributions in a dataset.

Retrievability of the metadata field *Data_Set_Distribution ->access_url; Distribution ->access_URL* (labelled as r50) is determined by performing a GET request, as shown in Code snippet 3.12. The GET request is sent to the URL written in the metadata field *Data_Set_Distribution ->access_url; Distribution ->access_URL* using httr package in R (Wickham, 2023a). To make the solution more robust, the request is sent two times within 5

https://github.com/opendatamonitor/odm.restapi/blob/master/odmapi/def_formatLists.py, https://gitlab.com/european-data-portal/edp-vocabularies/-/blob/master/edp-machine-readable-format.rdf

```
nonProprietaryMediaTypes <- read.csv("non_proprietary.csv")
numberOfMatches <- which(nonProprietaryMediaTypes$codes == tolower(resourceFormat))
if(length(numberOfMatches) > 0)
{
        resource_format_non_proprietary <- resource_format_non_proprietary + 1
}
```

Code snippet 3.10: Evaluation of the openness (open / non-proprietary formats) of the metadata field field *Data_Set_Distribution ->format; Distribution ->format* at the distribution level.

```
machineReadableMediaTypes <- read.csv("machine_readable.csv")
numberOfMatches <- which(machineReadableMediaTypes$codes == tolower(resourceFormat))
if(length(numberOfMatches) > 0)
{
        resource_format_machine_readable <- resource_format_machine_readable + 1
}
```

Code snippet 3.11: Evaluation of the openness (machine-readable file formats) of the metadata field *Data_Set_Distribution ->format; Distribution ->format* at the distribution level.

seconds in case the portal is unavailable at a given time. Then, the number of distributions with a retrievable value of the field *Data_Set_Distribution ->access_url; Distribution ->access_URL* of dataset is divided by the total number of distributions attached to a dataset.

The timely of the metadata field *Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* (labelled as c2) is determined by checking whether the value of the field *_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* of each distribution associated with the dataset is within the period defined in the field *Data_Set ->accrual_periodicity; Dataset ->frequency* (see Code snippet 3.13). For example, if the value of the metadata field *Data_Set ->accrual_periodicity; Dataset ->frequency* is ‘yearly’, it is checked whether the value of the metadata field *Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* of each dataset distribution is within one year. The R package chron is used to easily work with and manipulate dates (James & Hornik, 2023). Then, the number of distributions with a timely value of the field *Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* of a dataset is divided by the total number of distributions attached to a dataset.

Finally, the datasets from the two open data portals are combined into a single dataset for further analysis. This dataset contains values of individual indicators as variables (columns), with open datasets as observations (rows). The portal from which the dataset metadata are taken, and the URL of the open dataset are also included.

Therefore, the next step involves multivariate data analysis. The criteria for choosing the

```
tryCatch(
{
        response <- GET(accessUrl, timeout(timeToWaitForResponseInSeconds))
        if(response$status_code == 200)
        {
                resource_access_url_retrivable <- resource_access_url_retrivable + 1
        }
}
,error = function(err) {
        Sys.sleep(5)
        response <- GET(accessUrl, timeout(timeToWaitForResponseInSeconds))
        if(response$status_code == 200)
        {
                resource_access_url_retrivable <- resource_access_url_retrivable + 1
        }
})
```

Code snippet 3.12: Evaluation of the retrievability of the metadata field *Data_Set_Distribution ->access_url; Distribution ->access_URL* at the distribution level.

appropriate multivariate method for data analysis are the type of the data, the nature of the problem, and the goals to be achieved by the analysis (OECD et al., 2008; Chatfield & Collins, 1980). Regarding the type of data, multivariate data are mixed because they contain both qualitative and quantitative variables. As for the nature of the problem, the theoretical framework is defined as a formative model, which means that causality leads from the indicators to the construct and not vice versa. Considering the last criterion, multivariate analysis is used primarily for descriptive purposes. Based on those criteria, the hierarchical clustering, principal components analysis (PCA), and factor analysis of mixed data are selected and applied (Field et al., 2012; Gareth, Witten, Hastie, & Tibshirani, 2013; Chatfield & Collins, 1980). In addition, only the principal components analysis is described below, as its results have proven to be the most interpretable compared to other methods.

Before the principal components analysis is carried out, the prepared data are analysed using descriptive statistics. If a particular variable behaves like a constant, i.e., all observed open datasets receive the same value for an individual indicator, it should be removed from further consideration. If there are several variables that behave identically, it is sufficient to include only one in the further analysis.

To apply PCA, it should be determined whether it will be applied to the raw data (i.e., a dataframe that contains the scores of the individual indicators for the number of open datasets and from which, as already described, redundant variables are excluded) or to the correlation matrix. As stated in (Field et al., 2012), if the analysis is performed for at least 100,000 objects, the use of the correlation matrix is recommended, but in general it does not matter whether the

```
switch(tolower(row$dataset_accural_periodicity),
"decenial" = {
        if(currentDate-365*10 < date)
        {
                resource_accural_periodicity_timely
                <- resource_accural_periodicity_timely + 1
        }
},
"r/p10y" = {
        if(currentDate-365*10 < date)
        {
                resource_accural_periodicity_timely
                 <- resource_accural_periodicity_timely + 1
        }
},
"quadrennial"= {
        if(currentDate-365*4 < date)
        {
                resource_accural_periodicity_timely
                <- resource_accural_periodicity_timely + 1
        }
},
"r/p4y"= {
        if(currentDate-365*4 < date)
        {
                resource_accural_periodicity_timely
                <- resource_accural_periodicity_timely + 1
        }
},
"annual"= {
        if(currentDate-365 < date)
        {
                resource_accural_periodicity_timely
                <- resource_accural_periodicity_timely + 1
        }
},
...
```

Code snippet 3.13: Evaluation of the retrievability of the timeliness of the metadata field *Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* at the distribution level.

analysis is based on the correlation matrix or the raw data. The Pearson correlation coefficient is calculated between each pair of variables and the resulting matrix is plotted using a function from the R package corrplot (Wei & Simko, 2021). Considering that, in PCA, the number of components for extraction must be determined in advance, Horn's parallel analysis (Horn, 1965) was performed with the R package psych (Revelle, 2023), as was done in the Q methodology. When the loadings of the variables on each principal component have been computed by PCA, most of the variables have had high loadings on the first component — the component with the largest eigenvalue. This in turn can complicate the interpretation of the results obtained. For this reason, it is advisable to apply rotation techniques to improve the interpretability of the components (Field et al., 2012). The rotation/transformation that provided the best separation of components is the rotation 'cluster'. This rotation, as described by Revelle (2023), 'does a targeted rotation to a structure defined by the cluster representation of a varimax solution'. The function from the R package psych is used to perform PCA and the component rotation technique (Revelle, 2023). To assess the fit of the created principal components model, the differences between the observed correlations and the correlations based on the model, the so-called residuals, should be calculated (Field et al., 2012). In addition, there are several measures to define how small the residuals should be, and two of them are used in this research (Field et al., 2012):

1. Fit based on off-diagonal values – a value greater than 0.90 is an indicator of a good fit,

2. Root-mean-square residuals – value less than 0.08 is an indicator of a good fit.

The loadings of the variables on the principal components are visualised by drawing them as a network diagram using the functions of the R package qgraph (Epskamp, Costantini, Haslbeck, & Isvoranu, 2023). Furthermore, the results of the PCA are displayed graphically by combining a score plot (to visualise the data by projecting the observations onto the range of the two PCs) and a loadings plot (to visualise the data by projecting the variable vectors onto the range of the PCs) in a diagram, a so-called biplot. The biplots are created for the first three components, i.e., for those that explain the largest proportion of the variance in the data.

After a multivariate analysis, the normalization step is crucial for data comparability in the construction of composite indicators, and reasons for how and why the normalization should be performed are (OECD et al., 2008): the normalization procedure should be selected in a way that is appropriate to the theoretical framework and data properties; the way in which extreme values or outliers can become unintentional benchmarks should clarify their presence in the dataset; transforming highly distorted indicators if there is a need to do so. Given that the individual indicators are designed so that their score value can be expressed on a scale from 0 to 1, there was no need to apply the normalization procedure.

Individual indicators according to the methodology for constructing composite indicators should be weighted and aggregated with respect to the developed theoretical framework. Weighting can notably impact the overall values of a composite indicator (scores and ranks) assigned to open datasets when the composite indicator is used as a benchmarking framework. Although most composite indicators are based on the principle of equal weighting, this can lead to an unbalanced structure of the composite indicator, as the individual indicators are grouped into lower- and higher-level dimensions, and these into a composite indicator (OECD et al., 2008). Equal weighting at the level of the individual indicators could mean unequal weighting at the level of the dimensions, since the dimensions may contain more/fewer individual indicators. Also, if all individual indicators are given the same importance, there is a risk of double counting when highly correlated indicators are combined (OECD et al., 2008).

The analytic hierarchy process (AHP), a multi-criteria decision-making method, is used to determine the weights of the theoretical framework's elements in this research. The selection of this method is based on the characteristics of the problem. The process of determining weights involved two groups of experts. The first group included experts within the public administration with a focus on OD/IT strategists within the public sector, while the second one included experts outside the public administration with a focus on open data users, primarily researchers (Simperl et al., 2014). Only experts who helped in testing the content validity of the proposed theoretical framework (i.e., identifying which individual indicators are relevant for a particular quality dimension of open data metadata) are invited to participate in the weighting of the elements of the theoretical framework. An invitation for participation with detailed instructions was sent to the experts via e-mail, as shown in Appendix B. Additionally, the following documents are attached to the invitation email: Microsoft Excel document entitled 'The Importance of Elements in the Theoretical Framework', which served as a form for collecting expert opinions on the relative importance of dimensions as well as individual indicators; Microsoft Word document entitled 'Definitions of Dimensions', which contained a description of each dimension (as defined in subchapter 4.1.1, page 50); and a Microsoft Excel document entitled 'Descriptions of Metadata Fields', which contained the mapping of metadata fields and their descriptions of different metadata standards (as described in subchapter 2.2.3, page 23). Also, the email contained a link to the prepared video tutorial[14], which provided a guide on how to perform pairwise comparisons and maintain consistency in judgements.

Relative priorities are calculated based on pairwise comparisons of the theoretical framework's elements. These comparisons are conducted by experts who fill out the form (see Appendix C). The pairwise comparisons are made using a scale of absolute judgements that represent how much more one element dominates another with respect to a given property (Saaty, 1990, 1977). Experts should compare elements of the same level of the theoretical framework in pairs with respect to elements of the higher level. Therefore, individual indicators should be compared in pairs with respect to the dimension to which they belong. For instance, an expert should indicate how many times more, or how much stronger, individual indicator f20 is relevant for the findability dimension compared to individual indicator f21. Similarly, dimensions should be compared in pairs with respect to the composite indicator. Considering that even experts could not be completely consistent in their judgments, a consistency ratio (CR) should be calculated to determine the potential inconsistency of the judgments (Saaty, 1977). Saaty (1977) recommended that assessments should be accepted if CR < 0.10, which means that the expert providing their judgments was reasonably inconsistent.

[14] https://www.youtube.com/watch?v=IclCMKBeM38

Considering that the decision about the relative importance of the elements of the theoretical framework is not made by an individual, but by a group of individuals in which all individuals are considered equally important, it is necessary to summarise the group's judgements before calculating the relative priorities (Saaty & Begičević Ređep, 2012). As explained by Basak and Saaty (1993); Aczél and Saaty (1983), the geometric mean is the appropriate way for synthesizing the judgements, thus it is applied within this research. Subsequently, the relative priorities on the aggregated judgement matrix are determined using the eigenvector method (Saaty, 1977, 1990). Additionally, the consistency ratio of the aggregated judgements is calculated as suggested by Saaty (1977, 1990). The calculation of the relative priorities of the aggregated judgements and the consistency ratios is carried out in R. The packages readxl, tidyverse, xtable are used to load and analyse data, and to create tables (Wickham & Bryan, 2023; Wickham, 2023c; Dahl et al., 2019).

Individual indications are combined into a single value through the process of aggregation. Since many composite indicators have a hierarchical structure, multiple aggregations are needed. In other words, to reach the final composite value, sets of indicators should be combined into aggregate values of lower level, which should then be combined into higher-level aggregates. The complexity of the composite indicator structure, or the total number of levels in the hierarchical structure, determines how many lower and higher-level aggregations are required.

There are different methods of aggregation, and each has its advantages and disadvantages (OECD et al., 2008; Aoki, Kim, & Lee, 2013). While linear aggregation is useful when all individual indicators are expressed in the same measurement unit, geometric aggregation methods are more appropriate for a certain degree of non-compensation among individual indicators or dimensions (OECD et al., 2008). Given the method's broad application, the linear aggregation is used within this research since all the necessary conditions are met (e.g., normalised data, compatible with used weighting method).

The aggregation of the elements is done by constructing a composite indicator using functions from the COINr package in R (Becker, 2023). To calculate the composite scores for each open dataset, the entire structure had to be defined in advance (e.g., the aggregation level for each element in the structure, the parent–child relationships between the elements, the weighting of each element), as shown in the Code snippet 3.14.

3.2.4 Demonstrate artefact

Within this activity, the application of the artefact is demonstrated in a specific case, indicating its feasibility (Johannesson & Perjons, 2014; Peffers et al., 2007). Therefore, in order to determine whether the composite indicator addresses the identified problem, it is applied to a specific case (Prat, Comyn-Wattiau, & Akoka, 2015).

Descriptive statistics (minimum, maximum, mean, median, 1st and 3rd quartiles, standard deviation and interquartile range) of the dimension scores and the composite indicator, broken down by portal, are reported for the selected sample of datasets. The distributions of the di-

```r
library(COINr)
library(tidyverse)

load("../Data fetch and metrics calculation/mergedProcessedMetadata.Rda")
preparedDataFrame <- mergedProcessedMetadata[,-c(2,32)]

preparedDataFrame <-  cbind(preparedDataFrame, u_c3=0)
preparedDataFrame <-  cbind(preparedDataFrame, c9=0)
preparedDataFrame <-  cbind(preparedDataFrame, u59=0)

iCode <- c("f20","f21","f23","f24","f33","f34",
"i37","i38","i39","i40","i41","i42",
"u54","u55","u56","u58","u59","u61","u_c3",
"c2","c4","c9","c11","c12","c13","c18","c_f22","c_f31",
"r48","r49","r50","r51",
"F","R","I","U","C",
"CI")

Weight <- c(0.1500632,0.1283981,0.1217927,0.1370599,0.1987733,0.2639128,
0.1130808,0.2291885,0.2255983,0.1624630,0.1690836,0.1005858,
0.1210262,0.1198933,0.1437550,0.1660830,0.1266814,0.1649290,0.1576321,
0.07044559,0.06749004,0.06887624,0.09742347,0.11750508,0.07697634,0.12674593,
0.13719697,0.23734033,
0.2574494,0.1822675,0.1996726,0.3606105,
0.3145888,0.2342226,0.1658129,0.1172300,0.1681457,
1)

Level <- rep(c(1), each=32)
Level <- c(Level, rep(c(2),each=5))
Level <- c(Level, 3)

Parent <- c("F","F","F","F","F","F",
"I","I","I","I","I","I",
"U","U","U","U","U","U","U",
"C","C","C","C","C","C","C","C","C",
"R","R","R","R",
"CI","CI","CI","CI","CI",
NA)

Type <- rep(c("Indicator"), each=32)
Type <- c(Type, rep(c("Aggregate"),each=6))

metaDataFrame <- data.frame(iCode, Level, Parent, Type, Weight)
metaDataFrame <- cbind(metaDataFrame, Direction=1)

coin <- new_coin(preparedDataFrame, metaDataFrame)
coin <- Aggregate(coin, "Raw")
```

Code snippet 3.14: The process of building the composite indicator.

mension scores and the composite indicator are visualised using boxplots. The structure of the composite indicator for some individual open datasets is visualised as a radial plot using the R package fmsb (Nakazawa, 2024).

3.2.5 Evaluate artefact

The main objective of the activity Evaluate artefact is to determine how well the artefact meets the defined requirements and how much it can resolve or mitigate the real-world issue that prompted the research (Johannesson & Perjons, 2014). A combination of uncertainty analysis (UA) and sensitivity analysis (SA) is used to evaluate the robustness of the developed metadata quality composite indicator for open datasets. Uncertainty analysis is referred to as the process by which uncertainty in the inputs spreads through the model/system (here, the structure of the composite indicator) and affects the uncertainty in the outputs (here, the values of the composite indicator). Input uncertainties (assumptions, parameters) in the context of this research are referred to as actions taken during certain steps of constructing a composite indicator, for example, the inclusion or exclusion of individual indicators, the use of alternative methods for data normalisation, and the application of different weighting methods (Saisana et al., 2005; OECD et al., 2008). Sensitivity analysis is used to estimate which of the input uncertainties contribute to the output uncertainty and by how much. Through the application of sensitivity analysis, it can be recognised which input uncertainties have the most substantial influence on the model, i.e., the composite indicator (thus possibly deserving additional attention), distinguishing them from those that do not (Saisana et al., 2005; OECD et al., 2008). Utilising uncertainty and sensitivity analyses together can increase indicator transparency, test the robustness of composite indicator scores/ranks, and discover datasets that are scored/ranked better or worse due to certain assumptions (OECD et al., 2008). Assessing the robustness of composite indicators by the two analyses has proven useful in practice (OECD et al., 2008; Saisana et al., 2005).

Four uncertainty analyses and one sensitivity analysis are performed using the COINr package in R (Becker, 2023), which are described in detail below. In addition, the R package tidyverse together with other R packages included in it (e.g., ggplot2, dplyr, tibble), facilitated data import, tidying, manipulation and visualisation (Wickham, 2023c; Wickham, Chang, et al., 2023; Wickham, François, Henry, Müller, & Vaughan, 2023; Müller & Wickham, 2023). Conducted uncertainty analyses and a sensitivity analysis follow the approach of Saisana et al. (2005). Therefore, the analyses are performed on the basis of the Monte Carlo approach, which essentially consists of recalculating values of the composite indicator many times. Each time, input uncertainties/assumptions are randomly varied to estimate the output distributions.

To perform uncertainty and sensitivity analyses, it is necessary to determine which assumptions should be treated as uncertain and which alternative values should be assigned to each assumption. Thus, in the first uncertainty analysis (1st UA), two input uncertainties/assumptions are tested:

1. The weights – This refers to the overall perturbation of weights for elements at all levels

of the composite indicator. The function for changing the weights is designed to generate replications of the original set of weights, where some random noise should be added to each replication according to the specification (see Code snippet 3.15). It is implemented based on (Foster, McGillivray, & Seth, 2009; Permanyer, 2011; Wong, 1998). The weights, or more precisely the vectors of the weights, are generated as a convex combination of the original weights (with a coefficient of 0.95) and a perturbation from the Dirichlet distribution (with a coefficient of 0.05). The reason for using the Dirichlet distribution is that it generates vectors with non-negative elements whose sum is 1. In addition, a convex combination of two vectors with non-negative elements and a sum of 1 again produces a vector with non-negative elements whose sum is 1. This ensures that the generated weights are valid for elements of the developed composite indicator. It should be noted that in the R function rdirichlet() from the package MCMCpack (Martin, Quinn, & Park, 2022), the value of the shape parameter, is set to 1, as this reduces the probability that the random vector is close to a unit vector. This means that more than one weight will be perturbed simultaneously most of the time.

```r
make_weights1 <- function(w, factor=0.05){
        n <- length(w)
        w.out <- w*(1-factor) + factor*MCMCpack::rdirichlet(1, rep(1,n))
        w.out
}
```

Figure 3.15: Perturbation of weights with a coefficient of 0.05.

2. The aggregation methods - As described in the subchapter 3.2.3, linear aggregation is used for building a composite indicator. However, the geometric mean can also be considered as an alternative. Since an individual indicator can reach a value of 0, and the geometric mean cannot be applied to values of 0, the original geometric aggregation has been slightly modified, as shown in Code snippet 3.16.

```r
agg_mean_plusone <- function(x,w){
        a_gmean(x+1, w)-1
}
```

Figure 3.16: Adjusted geometric mean.

The behaviour of the composite indicator values (i.e., scores and ranks) in 2,000 simulations (i.e., replications or runs) in the 1st UA is visualised using the R package vioplot (Adler, Kelly, Elliott, & Adamson, 2022).

Considering that more than one assumption is tested only in the 1st uncertainty analysis, only one sensitivity analysis is performed to determine how the given composite indicator depends on the input information. In sensitivity analysis, the same assumptions are treated as uncertain, and the same alternative values are assigned to each assumption as in the case of the 1st UA. As described by OECD et al. (2008); Saisana et al. (2005), composite indicators can be considered as models. When multiple layers of uncertainty coexist, a composite indicator can turn into a non-linear, potentially non-additive model. It is also pointed out that, for non-linear models, robust model-free techniques, such as variance-based techniques, should be used for sensitivity analysis (OECD et al., 2008; Saltelli et al., 2008). The literature (OECD et al., 2008; Saisana et al., 2005; Saltelli et al., 2008), also emphasises that sensitivity measures/indices based on the decomposition of the variance of the model output have been shown to be useful. Furthermore, Saltelli et al. (2008) found that a good, synthetic representation of the sensitivity pattern in a model with k input uncertainties can be achieved by considering the full set of first-order indices together with the total effects indices. According to Saltelli et al. (2008), one of the two basic methods for computing sensitivity indices is the Monte Carlo-based design developed by Saltelli (2002), which is an improved approach of Sobol' (1993) and Homma and Saltelli (1996). This improved method requires N(k+2) model evaluations, where N is a base sample Saltelli et al. (2008). The described method is applied within this research.

It should be noted that COINr calculates sensitivity measures with respect to the average absolute rank change between original/unperturbed and perturbed values (Becker, 2023). The unperturbed composite indicator (originally built) is replicated 12,000 times. In addition, the parameter number of bootstrap samples to be taken in estimating confidence intervals for sensitivity indices is set to 1,000.

In the second uncertainty analysis (2nd UA), only one assumption is tested, namely the change in the weights of the elements at the lowest level of the developed composite indicator, i.e., the individual indicators. To clarify, the effect (what will happen) of removing individual indicators one at a time is being tested. Since no suitable existing R function was found that would set the weight of an individual indicator to 0 and simultaneously recalculate the other values so that the sum of the weights of the remaining indicators of a given dimension remains equal to 1, a new function was implemented, as shown in Code snippet 3.17. Given that the

```
make_weights2 <- function(w, i){
        w.out <- w
        w.out[i] <- 0
        w.out <- w.out/sum(w.out)
        w.out
}
```

Figure 3.17: Weighting of individual indicators when one is removed.

composite indicator developed contains 32 individual indicators, the function shown in Code snippet 3.17 had to be repeated 32 times to generate 32 sets of recalculated weights. On this basis, the composite indicator is replicated 32 times. Similarly to the 1st UA, the results of the 2nd UA are presented using violin plots (Adler et al., 2022).

The third and fourth uncertainty analyses are used not only to observe how an uncertain assumption spreads through the structure of the composite indicator and affects the values of the composite indicator (i.e., the scores and ranks), but also to assess robustness. In addition, assumptions, the choice of aggregation method (under the 3rd UA) and changes in the weights of the elements of the composite indicator at all levels (under the 4th UA) are tested separately.

Therefore, in the 3rd uncertainty analysis, the robustness of the composite indicator (in two outputs assigned to each observed dataset, composite indicator scores and ranks) is observed with respect to the chosen aggregation method. Two alternative values (linear and geometric aggregation) are assigned to this one input assumption (the aggregation methods) and the composite indicator is replicated 4 times. In addition, the Pearson's correlation coefficient between the values of the composite indicator (i.e., scores and ranks) from different simulations is calculated, using the linear aggregation in one simulation and the geometric aggregation method in the other. The correlation is computed using the R package DescTools (Signorell, 2023) and visualised using the R package ggplot2 (Wickham, Chang, et al., 2023).

The choice of the aggregation method is not a significant source of uncertainty, if the Pearson correlation between the scores (or ranks, respectively) from two simulations using different aggregation methods is large enough. A one-tailed z-test, based on Fisher's z-transformation (Field et al., 2012; Zimmerman, Zumbo, & Williams, 2003), at the statistical significance level $p < 0.05$, is performed to test the hypotheses:

- Null hypothesis: The correlation between the scores or ranks of the composite indicator obtained by different aggregation methods is equal to 0.99.
- Alternative hypothesis: The correlation between the scores or ranks of the composite indicator obtained by different aggregation methods is greater than 0.99.

The function implementing the z-test is shown in Code Snippet 3.18. In the 4th uncertainty

```r
my_ztest <- function(rho, rho0=0, n){
        z <- atanh(rho)
        z0 <- atanh(rho0)
        zscore <- (z-z0)/sqrt(1.06/(n-3))
        p <- 1-pnorm(zscore)
        return(p)
}
```

Figure 3.18: Z-statistics based on Fisher's z-transformation.

analysis, the robustness of the composite indicator in an output, i.e., the open dataset ranking, is observed with respect to the overall perturbation of the weights of the elements at all levels of the composite indicator. In addition to the use of functions from the already mentioned R packages COINr, tidyverse, MCMCpack, functions from the packages doParallel and foreach are also used (Becker, 2023; Wickham, 2023c; Martin et al., 2022; Microsoft Corporation & Weston, 2022a, 2022b). Functions from these packages are used to perform computations in parallel and thus reduce the computing time. Perturbations of weights are done in a similar way as in the 1st UA. In this analysis, the coefficient of the random component ranges between 0.01 and 0.95 (30 different values). The value of the Dirichlet distribution shape parameter is set at 0.01, which means that the probability of a random vector close to a unit vector is higher than in the 1st UA (see Code snippet 3.19). For each value of the coefficient of the random component, 2,000 simulations are run. If the composite indicator is robust to a perturbation in weights, it will not change the order of a pair of datasets in the ranking. For each value of the perturbation coefficient, the number of dataset pairs that change their order in at least one of the 2,000 simulations is counted. Tied ranking is not counted as a change.

```r
make_weights3 <- function(w, factor=factors){
        n <- length(w)
        w.out <- w*(1-factor) + factor*MCMCpack::rdirichlet(1, rep(0.01,n))
        w.out
}
```

Figure 3.19: Perturbation of weights with various coefficient values.

CHAPTER 4

RESULTS

This chapter contains the results of this research, which is conducted through a number of related activities and steps as described in the previous chapter. They are presented in the form of a theoretical framework of metadata quality for open datasets (see subchapter 4.1) and a composite indicator of metadata quality of open datasets (see subchapter 4.2).

4.1 Theoretical framework of metadata quality for open datasets

This subchapter focuses on the development of the theoretical framework for the metadata quality of open datasets. First, the development of the initial version of the framework is presented, from defining the structure itself to identifying elements at each level. Then, the results of the process of identifying individual indicators relevant to a particular quality dimension of open data metadata based on expert opinion are described and the final version of the theoretical framework is defined. Furthermore, the theoretical framework is developed following two activities of the method framework for design science research and two steps of the methodology for constructing composite indicators. Therefore, this subchapter presents the results obtained, which are directly related to the development of the theoretical framework for the metadata quality of open datasets, following the first two main activities of the method framework for design science research, i.e., explicating the problem and defining requirements, as the method framework for design science research represents the outer cycle of conducting this research.

4.1.1 Explicate problem

The trigger for this activity is the existence of a problem, namely the questionable quality of open (meta)data, which is relevant not only for a local but also for a global practise, as explained in more detail in subchapter 1.1 (page 4). To contribute to solving the noticed problem, it is crucial to understand the concept of quality as well as other concepts related to open data, metadata and the measurement of metadata quality. Therefore, the concepts important to this research are described and defined in chapter 1 and chapter 2. A catalogue of common concepts (see Glossary) is also created.

In order to improve on previous efforts to solve the practical problem described, the main objective of this research was to develop a theoretical framework of open metadata quality and operationalise it through a new composite indicator that allows the comparison of metadata of open datasets. Therefore, the following criteria were defined for the development of a composite indicator:

- The composite indicator should be constructed based on metadata about open datasets that are automatically retrieved from portals without human intervention;

- The composite indicator should facilitate comparisons within a portal, between different portals and at different points in time, which may lead to the exclusion of some individual indicators identified in the literature;
- An individual indicator that depends on the data/metadata of a specific portal is not appropriate;
- Applying the same individual indicator to the same data should always result in the same value.

Defining the structure and all elements of the theoretical framework represents the starting point in constructing a composite indicator. The structure itself is roughly outlined by the methodology for constructing composite indicators as described in subchapter 3.2.1 (page 27). At the lowest level, individual indicators are positioned, and then aggregated into higher-level elements, i.e. the dimensions. These dimensions are further aggregated into a meaningful composite indicator under a fitness-for-purpose principle. The elements of the theoretical framework were selected, defined and combined into higher-level elements on the basis of a review of the literature (Šlibar et al., 2018, 2021). As the systematic literature review referred only to academic papers assessing the quality of open data (Šlibar et al., 2021) and not to previous efforts by practitioners, all such documents, reports and other sources were additionally reviewed. The literature review revealed that practitioners and academics have different views on the assessment of the quality of metadata in open datasets. Practitioners assess the metadata quality of open data based on the properties of the dataset that are supported by the metadata (e.g., findability, interoperability). In contrast, the academic community evaluates the quality of metadata primarily based on the properties of the metadata itself (e.g. completeness, accuracy). To ensure that neither perspective is overlooked, the dimensions of the theoretical framework are defined in terms of the properties of the dataset and in terms of the properties of the metadata.

In the construction of the hierarchical structure of the theoretical framework, the metadata fields that were found to be relevant according to the metadata standards formed the central point from which the other elements of the theoretical framework were defined (Šlibar, 2024b). Therefore, the dimensions of both perspectives are merged on the basis of metadata fields and associated metrics:

- If a particular metadata field was not covered in the existing dimensions in terms of metadata properties, it was checked whether it was covered in the existing dimensions in terms of dataset properties. If it was covered by the existing dimensions in terms of dataset properties, the name of the dimension was defined based on the metric associated with the metadata field.
- If a particular metadata field was not covered in the existing dimensions in terms of dataset properties, it was first checked whether a metric was associated with the observed

metadata field in terms of metadata properties. If the metadata field was covered in the existing dimensions related to the metadata properties, it was assigned the existing dimension related to the dataset properties based on other metadata fields that comprise that dimension in terms of dataset properties.

- If a particular metadata field was not covered in either the existing dimensions related to the metadata properties or the existing dimensions related to the dataset properties, it was assigned dimensions from both perspectives based on other metadata fields comprising the dimensions.

Looking at the properties of the metadata itself, a total of eight dimensions are defined (Šlibar et al., 2021; Reiche & Höfig, 2013; Király, 2019; Neumaier et al., 2016):

- **Accuracy** assesses the precision of the information contained in the metadata. It can also be defined as the correspondence of metadata to actual data (more specifically, to the resources) or to quality certification document or similar documents.
- **Coherence** measures the degree to which all metadata uniformly describe a particular object.
- **Completeness** refers to the extent of the information present in the metadata.
- **Conformance** means the absence of contradictions and reflects the logical consistency of the metadata with its preceding values, established norms, standards and other relevant criteria.
- **Openness** refers to the extent to which data/content (e.g., metadata, resources) conforms to open licences, are non- proprietary, and are machine-readable.
- **Retrievability** measures the success of fetching data/content (e.g., metadata, resources) by an agent.
- **Timeliness** indicates how often the data/content (e.g., metadata, resources) are updated to ensure relevance and currency.
- **Understandability** refers to the clarity and comprehensibility of the information contained in the metadata.

Taking into account the properties of the dataset supported by the metadata, a total of five dimensions are defined (Consortium of data.europa.eu, n.d.; Deutz et al., 2020; Šlibar et al., 2021; Association for Computing Machinery, 2020): findability (F), retrievability (R), interoperability (I), reusability (U), and contextuality (C).

Findability

Findability is the extent to which humans and machines can easily discover (meta)data. In addition to information that helps both humans and machines to identify data uniquely and unambiguously, information about the temporal and geographic area(s) covered by the data is also relevant for this dimension (Consortium of data.europa.eu, n.d.; Deutz et al., 2020). For this dimension, 17 individual indicators were defined on the basis of the relevant literature (see Table 4.1).

Retrievability

Retrievability is the extent to which humans and machines can fetch (meta)data successfully (Consortium of data.europa.eu, n.d.; Deutz et al., 2020; Šlibar et al., 2021). For this dimension, six individual indicators were defined on the basis of the relevant literature (see Table 4.2).

Interoperability

Interoperability is the extent to which different applications and systems can successfully communicate and exchange data with unambiguous, shared meaning. Interoperability implies both syntactic interoperability (compatible formats and protocols) and semantic interoperability (uniform codification of data) (Consortium of data.europa.eu, n.d.; Deutz et al., 2020). For this dimension, 11 individual indicators were defined on the basis of the relevant literature (see Table 4.3).

Reusability

Reusability is the extent to which (meta)data are well-described so that data can be replicated by different teams within different experimental setups. The information about the terms and conditions on how the data could be accessed and reused is relevant for this dimension, as well as provenance information related to the data creation process, and who can be contacted for more information about the data (Consortium of data.europa.eu, n.d.; Deutz et al., 2020; Association for Computing Machinery, 2020). For this dimension, 18 individual indicators were defined on the basis of the relevant literature (see Table 4.4).

Contextuality

Contextuality is the extent to which the user can obtain additional information about the data (e.g., origin, quality, copyright statements, date of publication) (Consortium of data.europa.eu, n.d.). For this dimension, 19 individual indicators were defined on the basis of the relevant literature (see Table 4.5).

The hierarchical structure of the concepts, with the composite indicator at the top, the dimensions in terms of the properties of the dataset and in terms of the properties of the metadata below, and the individual indicators at the bottom of this structure, is shown in Figure 4.1. It

Table 4.1: The individual indicators of the findability dimension in the initial version of the theoretical framework.

Coherence	
f30	Semantic distance between the Designation [Data_Set_Distribution] ->sign; Distribution ->title value for each distribution attached to the dataset AND the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset
f32	Semantic distance between the Designation [Data_Set] ->sign; Cataloged_Resource ->title value AND the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value
Completeness	
f20	Existence of the Classification [Data_Set] ->classification_scheme_item_value; Cataloged_Resource ->keyword / tag value
f21	Existence of the Classification [Data_Set] ->classification_scheme_name; Cataloged_Resource ->theme / category value
f22	Existence of the Data_Set ->spatial_coverage; Dataset ->spatial / geographical_coverage value
f23	Existence of the Data_Set ->temporal_coverage_end_date; Dataset ->temporal_coverage value
f24	Existence of the Data_Set ->temporal_coverage_start_date; Dataset ->temporal_coverage value
f25	Existence of the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value
f31	Existence of the Designation [Data_Set] ->sign; Cataloged_Resource ->title value
f33	Existence of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value
f35	Existence of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset
Conformance	
f34	Conformity of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier
f36	Conformity of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier
Understandability	
f26	Readability of the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is computed by using the Flesch-Kincaid Reading Ease test
f27	Intrinsic precision of the text provided within Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is determined by spelling mistakes
f28	Readability of the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is computed by using the Flesch-Kincaid Reading Ease test
f29	Intrinsic precision of the text provided within Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is determined by spelling mistakes

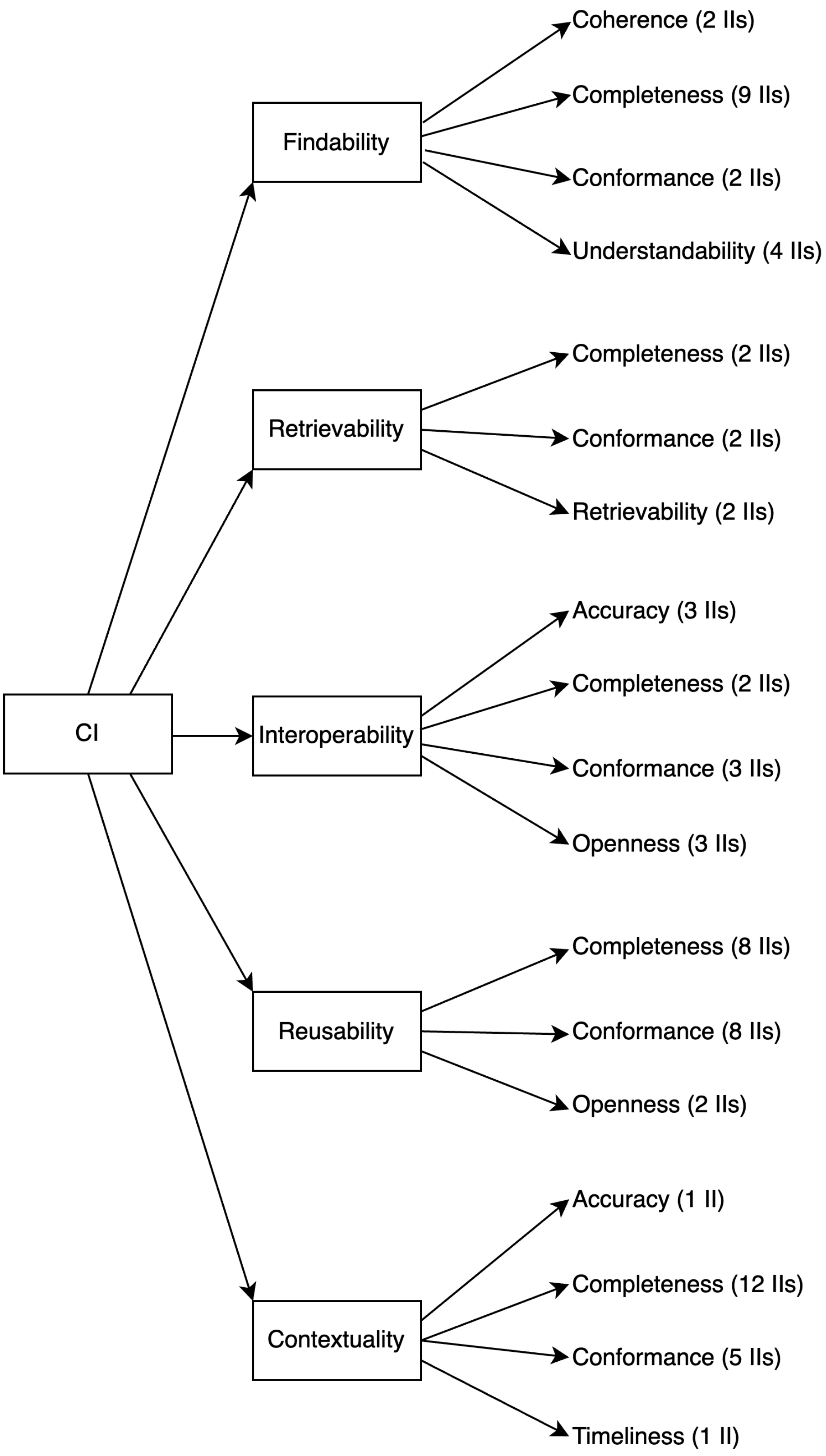


Figure 4.1: The initial theoretical structure of the composite indicator (Legend: The numbers in brackets indicate the number of individual indicators contained in the dimensions). The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59.

Table 4.2: The individual indicators of the retrievability dimension in the initial version of the theoretical framework.

Completeness	
r48	Existence of the Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset
r51	Existence of the Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset
Conformance	
r49	Validity of format of the HTTP URL provided within Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset
r52	Validity of format of the HTTP URL provided within Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset
Retrievability	
r50	Retrievability of the HTTP URL provided within Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset is determined based on an HTTP GET operation
r53	Retrievability of the HTTP URL provided within Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset is determined based on an HTTP GET operation

contains a total of five dimensions in terms of the properties of the dataset, eight in terms of the properties of the metadata and 71 individual indicators that have remained after the removal of semantically similar indicators. The initial version of the theoretical framework is stored as a dataset on the Harvard Dataverse research data repository (Šlibar, 2024c).

4.1.2 Define requirements

Since the Q methodology, in combination with Lawshe's content validity ratio, was chosen for collecting expert opinions on the initial theoretical framework and testing its content validity, a separate Q sorting task needed to be prepared for each dimension. For each Q sorting, the Q grid had to be determined based on the Q sample and the maximum Q sort rank:

- For findability – the Q sample contained 17 individual indicators that belong to the dimension and 16 random individual indicators that do not belong to it according to the literature (i.e., that are part of other dimensions); the maximum Q sort rank was set to four.
- For retrievability - the Q sample contained six individual indicators that belong to the dimension and six random individual indicators that do not belong to it according to the literature (i.e., that are part of other dimensions); the maximum Q sort rank was set to two.
- For interoperability - the Q sample contained 11 individual indicators that belong to the dimension and 10 random individual indicators that do not belong to it according to the literature (i.e., that are part of other dimensions); the maximum Q sort rank was set to three.

Table 4.3: The individual indicators of the interoperability dimension in the initial version of the theoretical framework.

Accuracy	
i38	Accuracy of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is computed by using file-extension of the actual resource and/or by taking the format specified in the HTTP content-type header field
i43	Accuracy of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-type header field
i46	Accuracy of the Registration_Authority [Data_Set] ->documentation_language_identifier; Cataloged_Resource ->language value is computed by using language detection on the actual resource and/or HTTP content-language header field
Completeness	
i37	Existence of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset
i42	Existence of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset
Conformance	
i39	Conformity of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset with one of the IANA media types
i44	Conformity of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset with one of the IANA media types
i47	Conformity of the Registration_Authority [Data_Set] ->documentation_language_identifier; Cataloged_Resource ->language value to a given standard such as ISO 639
Openness	
i40	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed open/non-proprietary formats
i41	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed machine-readable file formats
i45	Openness of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is checked based on a predefined set of confirmed open/non-proprietary formats

Table 4.4: The individual indicators of the reusability dimension in the initial version of the theoretical framework.

Completeness	
u54	Existence of the Data_Set ->rights; Cataloged_Resource ->access_rights value
u56	Existence of the Data_Set ->rights; Cataloged_Resource ->license value
u59	Existence of the Data_Set_Distribution ->rights; Distribution ->access_rights value for each distribution attached to the dataset
u61	Existence of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset
u64	Existence of the Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or the Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset
u67	Existence of the Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value
u70	Existence of the Stewardship_Record [Data_Set] ->organization; Organization / Person | foaf:Organization ->foaf:name value
u71	Existence of the Data_Set_Provenance ->generation_type; Dataset ->was_generated_by value
Conformance	
u55	Conformity of the Data_Set ->rights; Cataloged_Resource ->access_rights value with the EU controlled vocabulary for access rights
u57	Conformity of the Data_Set ->rights; Cataloged_Resource ->license value with one of the licenses from the predefined list provided by the Open Definition or the EU Vocabularies related to licenses
u60	Conformity of the Data_Set_Distribution ->rights; Distribution ->access_rights value for each distribution attached to the dataset with the EU controlled vocabulary for access rights
u62	Conformity of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset with one of the licenses from the predefined list provided by the Open Definition or the EU Vocabularies related to licenses
u65	Validity of format of the email address provided within Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset
u66	Validity of format of the HTTP URL provided within Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset
u68	Validity of format of the email address provided within Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value
u69	Validity of format of the HTTP URL provided within Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value
Openness	
u58	Openness of the Data_Set ->rights; Cataloged_Resource ->license value is checked based on the assessment of the Open Definition
u63	Openness of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset is checked based on the assessment of the Open Definition

Table 4.5: The individual indicators of the contextuality dimension in the initial version of the theoretical framework.

Accuracy	
c8	Accuracy of the Data_Set_Distribution ->size; Distribution ->byteSize value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-length header field
Completeness	
c1	Existence of the Data_Set ->accrual_periodicity; Dataset ->frequency value
c3	Existence of the Data_Set ->rights; Cataloged_Resource ->rights value
c4	Existence of the Data_Set_Distribution ->issued_date; Distribution ->release_date value for each distribution attached to the dataset
c6	Existence of the Data_Set_Distribution ->rights; Distribution ->rights value for each distribution attached to the dataset
c7	Existence of the Data_Set_Distribution ->size; Distribution ->byteSize value for each distribution attached to the dataset
c9	Existence of the Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime value
c11	Existence of the Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date value
c13	Existence of the Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value
c15	Existence of the NA; Distribution ->update / modification_date value for each distribution attached to the dataset
c17	Existence of the Registration_State [Data_Set] ->effective_date; Cataloged_Resource ->release_date value
c18	Existence of the Data_Set_Provenance ->originator; Cataloged_Resource ->creator value
c19	Existence of the Data_Set_Quality_Assessment ->statement; dqv:QualityAnnotation ->oa:hasBody value
Conformance	
c5	Validity of format of the date provided within the Data_Set_Distribution ->issued_date; Distribution ->release_date value for each distribution attached to the dataset (e.g., according to ISO 8601)
c10	Validity of format of the date provided within Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime value (e.g., according to ISO 8601)
c12	Validity of format of the date provided within Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date value (e.g., according to ISO 8601)
c14	Validity of format of the date provided within Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value (e.g., according to ISO 8601)
c16	Validity of format of the date provided within NA; Distribution ->update / modification_date value for each distribution attached to the dataset (e.g., according to ISO 8601)
Timeliness	
c2	The timely of the Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value is determined in relation to the Data_Set ->accrual_periodicity; Dataset ->frequency value

- For reusability - the Q sample contained 18 individual indicators that belong to the dimension and 17 random individual indicators that do not belong to it according to the literature (i.e., that are part of other dimensions); the maximum Q sort rank was set to four.
- For contextuality - the Q sample contained 19 individual indicators that belong to the dimension and 19 random individual indicators that do not belong to it according to the literature (i.e., that are part of other dimensions); the maximum Q sort rank was set to five.

Of the 64 experts who were invited to participate in testing the content validity of the initial theoretical framework, or more precisely, to identify which individual indicators are relevant for a particular quality dimension of open data metadata, 11 responded. All 11 experts provided their answers for the Q sorting designed for the quality dimension findability, while nine of them provided answers for the other dimensions - retrievability, interoperability, reusability, and contextuality.

Average score and Lawshe's content validity ratio are calculated for each individual indicator of the initial theoretical framework, as shown in Appendix D. Based on the criteria listed in subchapter 3.2.2 (page 28), individual indicators listed in Table 4.6 are selected as relevant ones. Eight individual indicators (i37, i40, c11, c18, f21, f22, f31, r51) proved to be relevant in two or more dimensions. In general, they have a higher mean Q sort rank in the dimensions in which they are located according to the initial theoretical framework. The individual indicators f22 and f31 changed the dimension in which they are located according to the initial theoretical framework. Although the individual indicator c3 is not recognised as relevant in the dimension in which it is located according to the initial theoretical framework (contextuality), it is recognised as relevant in the reusability dimension. This means that indicator c3 has also changed dimension. The final theoretical structure of the composite indicator of metadata quality of open datasets is shown in Figure 4.2.

The Q methodology analysis is used to better understand potential differences in the understanding of a dimension among experts. The quantitative aspect involves the use of factor analytic techniques, particularly principal components analysis, as a means of grouping like-minded individuals. In order to determine whether one or more perception profiles exist in relation to the importance of combinations of individual indicators of a particular dimension, the number of components had to be determined, as described in subchapter 3.2.2 (page 28). Horn's parallel analysis revealed that retaining one principal component is sufficient for two dimensions (findability and retrievability), while the other three dimensions (contextuality, interoperability and reusability) require the retention of two components. This indicates that only one expert profile was identified for findability and retrievability in terms of expert opinion/perception of the relevance of individual indicators. In the cases where two profiles are identified (contextuality, interoperability and reusability), it is found that the degree of agreement within each profile in terms of the number of experts is not high enough to identify relevant individual indicators. For this reason, the selection of indicators for both expert profiles in these dimensions is made

Table 4.6: Mean Q sort rank ($\bar{x}$) and Lawshe's content validity ratio (CVR) for individual indicators selected as relevant for each dimension (Legend: U - Reusability, I - Interoperability, C - Contextuality, F - Findability, R - Retrievability). The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59.

		u56	u58	u54	u61	u59	i40	u55	c3	c11		
U	$\bar{x}$	0.556	0.500	0.472	0.444	0.333	**0.250**	0.222	**0.194**	**0.083**		
	CVR	0.778	0.778	0.556	0.778	0.778	0.778	0.556	0.778	0.778		
		i41	i40	i42	r51	i37	i39	i38				
I	$\bar{x}$	0.370	**0.296**	0.296	**0.296**	**0.259**	0.222	0.111				
	CVR	1.000	0.556	1.000	0.333	0.778	0.556	0.778				
		f22	c18	c2	c11	c13	f31	c9	c4	f21	c12	
C	$\bar{x}$	**0.511**	**0.467**	0.444	**0.444**	0.400	**0.356**	0.311	0.244	**0.244**	0.111	
	CVR	1.000	0.778	1.000	1.000	0.556	0.556	0.778	0.778	0.778	0.556	
		f20	f21	f33	f31	f34	c18	f22	f24	i37	f23	i40
F	$\bar{x}$	0.545	**0.364**	0.341	**0.273**	0.250	**0.250**	**0.227**	0.205	**0.205**	0.182	**0.159**
	CVR	0.818	0.636	0.636	0.455	0.818	0.455	0.636	0.636	0.818	0.636	0.818
		r48	r50	r49	r51							
R	$\bar{x}$	0.667	0.500	0.389	**0.333**							
	CVR	1.000	1.000	1.000	1.000							

Note: An individual indicator that has a high $\bar{x}$ / CVR in two or more dimensions, or in a dimension other than the one in which it is located according to the initial theoretical framework, is highlighted with the same colour. Each indicator is left in only one dimension, namely the one in which it has the highest $\bar{x}$, which is indicated by the green colour of the $\bar{x}$.

as a consensus.

The theoretical framework, which contains the individual indicators selected as relevant, is a main output of the first two activities of the method framework for design science research and the first two steps of the methodology for constructing composite indicators. At the same time, it represents the input for the following activities and related steps.

4.2 Composite indicator of metadata quality of open datasets

This subchapter shows how the developed theoretical framework of open metadata quality is operationalised into a new composite indicator for solving problems related to the quality assessment of open metadata. Furthermore, the developed theoretical framework is operationalised through the remaining activities of the method framework for design science research and the steps of the methodology for constructing composite indicators.

In this subchapter, the results directly related to the composite indicator are presented. These are structured in alignment with the remaining main activities of the design science research method framework: Design and develop artefact, Demonstrate artefact, and Evaluate artefact.

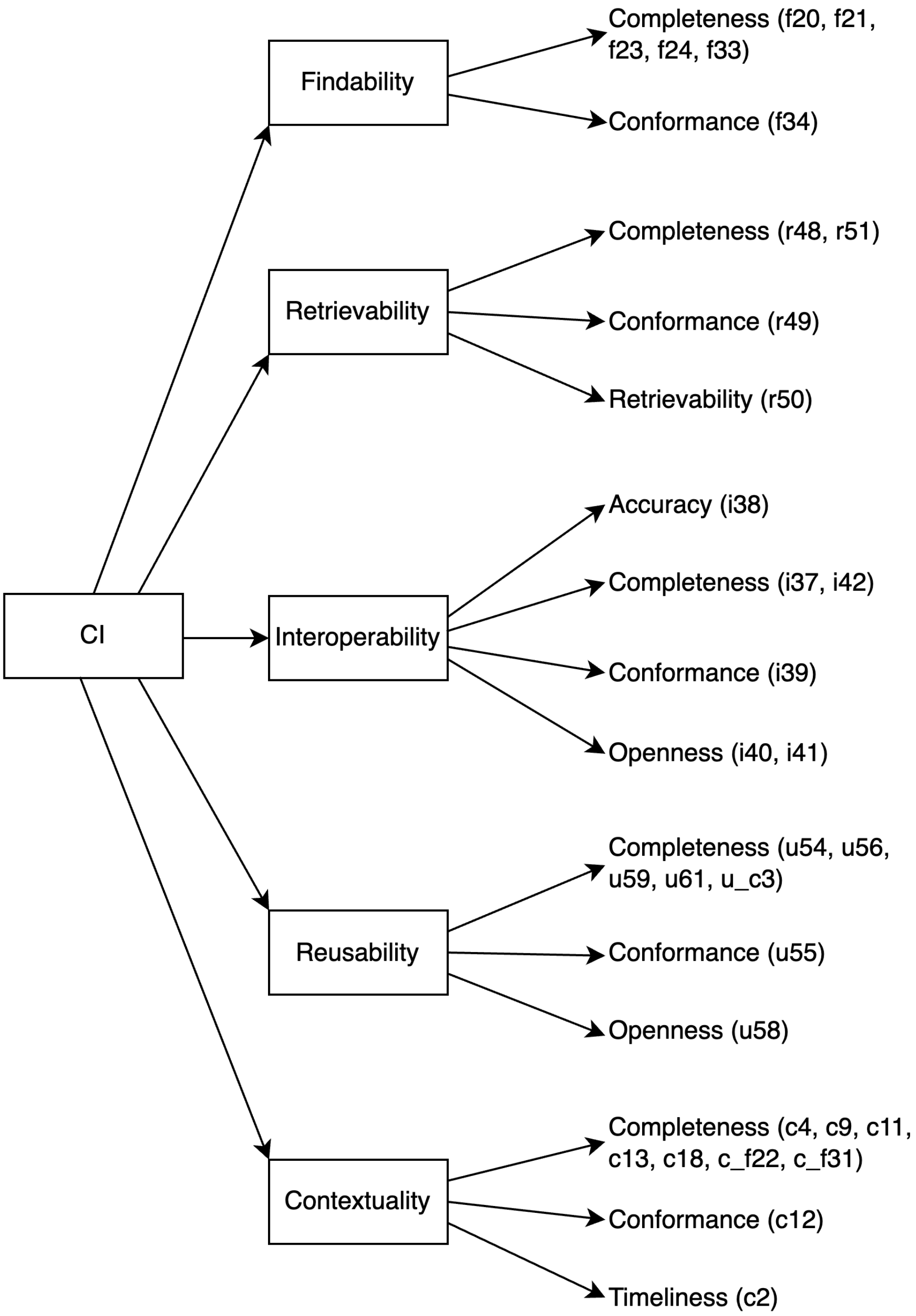


Figure 4.2: The final theoretical structure of the composite indicator. The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

4.2.1 Design and develop artefact

The activity of designing and developing an artefact began with the selection of variables (as described in subchapter 3.2.3, page 31), which includes the following:

1. Mapping the metadata fields of each selected OGD portal with those contained in (Šlibar, 2024b). However, the mapping is only performed for metadata fields that are used for the calculation of individual indicators of the developed theoretical framework. All individual indicators are based on one metadata field, with the exception of c2, which is based on two metadata fields. Table 4.7 indicates whether metadata field(s) of an individual indicator are found in real metadata and successfully mapped.

2. Calculating scores of the relevant individual indicators for each retrieved dataset. If a metadata field that is required for a relevant individual indicator is not found in the real metadata, the value of the individual indicator is set to 0. Therefore, some individual indicators are calculated for both OGD portals, some for one and some for neither portal. Of the total of 32 individual indicators, the metadata fields for three individual indicators u_c3, c9 and u59 (in particular the semantic mapping of the properties of the metadata fields in the real metadata to those defined in the relevant metadata standards) are not found on any of the OGD portals.

3. Combining all calculated scores of the individual indicators for the retrieved open datasets into a single dataframe, which is then stored as a dataset on the Harvard Dataverse research data repository (Šlibar, 2024a).

Table 4.8 shows descriptive statistics for dichotomous individual indicators. It can be observed that several individual indicators show extremely low variability. In particular, all, or almost all, observed open datasets achieve a value of 1 for the individual indicators f33, c11, and c12. On the other hand, all, or almost all, datasets achieve a value of 0 for the individual indicators c18, u54, u55, and, u58. This means that these individual indicators do little to distinguish the observed open datasets from each other. However, all of these individual indicators are not excluded from the composite indicator, as the situation on other portals could be completely different.

Furthermore, Table 4.9 shows descriptive statistics for continuous individual indicators. It can be observed that datasets across all quantitative continuous individual indicators reach the lowest value of 0 and the highest value of 1. Since the mean value for the individual indicators i37, r51 and, u61 is very high, namely 0.930, 0.901, and 0.877 respectively, it can be said that the vast majority of datasets achieve a very high, i.e., the best quality according to these indicators. This can also be seen in the values of the first quartile for the individual indicators mentioned (i37, r51, and u61), which are equal to 1, meaning that more than 75% of the datasets achieve the best quality for these individual indicators. On the other hand, the values of the third quartile for the individual indicators c2 and i42 are 0, which means that 75% of the datasets for these indicators achieve a value of 0, i.e., they are of the lowest quality. The individual indicator c2 has the lowest standard deviation compared to the other indicators, namely 0.06, which indicates that the variability between the datasets according to c2 is not high, i.e., they are all close to the mean. Although the indicators r48, r49, and u61 are described quantitatively, the datasets

Table 4.7: Existence of metadata field(s) required for the calculation of an individual indicator in real metadata. The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

Individual indicator	**European OGD portal**	**Australian OGD portal**
f20	✓	✓
f21	✓	✓
f23	✓	✓
f24	✓	✓
f33	✓	✓
f34	✓	✓
i37	✓	✓
i38	✓	✓
i39	✓	✓
i40	✓	✓
i41	✓	✓
i42	✗	✓
u54	✓	✓
u55	✓	✓
u56	✓	✗
u58	✓	✗
u59	✗	✗
u61	✓	✓
u_c3	✗	✗
c2	✗	✓
c4	✓	✓
c9	✗	✗
c11	✓	✓
c12	✓	✓
c13	✓	✓
c18	✓	✗
c_f22	✓	✓
c_f31	✓	✓
r48	✗	✓
r49	✗	✓
r50	✗	✓
r51	✓	✓

according to these indicators essentially only reach two different values (0 and 1). In contrast, datasets corresponding to the individual indicator i40 achieve up to 58 different values.

The descriptive data analysis is followed by the multivariate exploratory data analysis. Individual indicators for which there is no variability across datasets, i.e., those that are constant, are not included in the multivariate data analysis, namely c18, f33, u54, u55. In addition, individual

Table 4.8: Number and percentage of open datasets for which dichotomous individual indicators equal 1. Total number of open datasets (N) = 4622. The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

	N	p
c11	4578	(99.0%)
c12	4578	(99.0%)
c18	0	(0.0%)
c_f22	1435	(31.0%)
c_f31	2915	(63.1%)
f20	4286	(92.7%)
f21	1427	(30.9%)
f23	749	(16.2%)
f24	1051	(22.7%)
f33	4622	(100.0%)
f34	4231	(91.5%)
u54	0	(0.0%)
u55	0	(0.0%)
u56	1905	(41.2%)
u58	7	(0.2%)

Table 4.9: Descriptive statistics of continuous individual indicators (Legend: Min. - Minimum, Q1 - First quartile, $\bar{x}$ - Mean, Q3 - Third quartile, Max. - Maximum, SD - Standard deviation, IQR - Interquartile range, N_U - Number of unique values). The definitions of abbreviations for individual indicators are provided in Tables 4.2, 4.3, 4.4, 4.5, pages 56-59.

	Min.	*Q*1	*Median*	$\bar{x}$	*Q*3	*Max.*	*SD*	*IQR*	N_U
c13	0.000	0.000	0.000	0.265	0.904	1.000	0.435	0.904	29
c2	0.000	0.000	0.000	0.005	0.000	1.000	0.060	0.000	12
c4	0.000	0.000	0.000	0.477	1.000	1.000	0.499	1.000	3
i37	0.000	1.000	1.000	0.930	1.000	1.000	0.245	0.000	23
i38	0.000	0.000	0.000	0.355	1.000	1.000	0.444	1.000	57
i39	0.000	0.000	0.862	0.623	1.000	1.000	0.437	1.000	55
i40	0.000	0.000	0.108	0.384	1.000	1.000	0.438	1.000	58
i41	0.000	0.000	0.000	0.337	0.833	1.000	0.424	0.833	53
i42	0.000	0.000	0.000	0.159	0.000	1.000	0.352	0.000	49
r48	0.000	0.000	0.000	0.486	1.000	1.000	0.500	1.000	2
r49	0.000	0.000	0.000	0.486	1.000	1.000	0.500	1.000	2
r50	0.000	0.000	0.000	0.408	1.000	1.000	0.491	1.000	9
r51	0.000	1.000	1.000	0.901	1.000	1.000	0.299	0.000	6
u61	0.000	1.000	1.000	0.877	1.000	1.000	0.329	0.000	2

indicators c12 and r49 are excluded because c12 is identical to c11 for all observed open datasets, while r49 has the same value as r48 for all observed open datasets, thus making them redundant for multivariate analysis.

The correlations between the remaining individual indicators are calculated and plotted in the form of a correlation matrix, which serves as input for the principal components analysis (see Figure 4.3). In the figure, two blocks of individual indicators are visible. In the top left corner (first block) and in the bottom right corner (second block) of the correlation matrix there are indicators that are more or less positively correlated with each other. In contrast, the top right and bottom left corners show that the correlations between the individual indicators in the first block and the individual indicators in the second block are negative or very low. It can also be seen that the individual indicator u58 is not correlated with the other individual indicators.

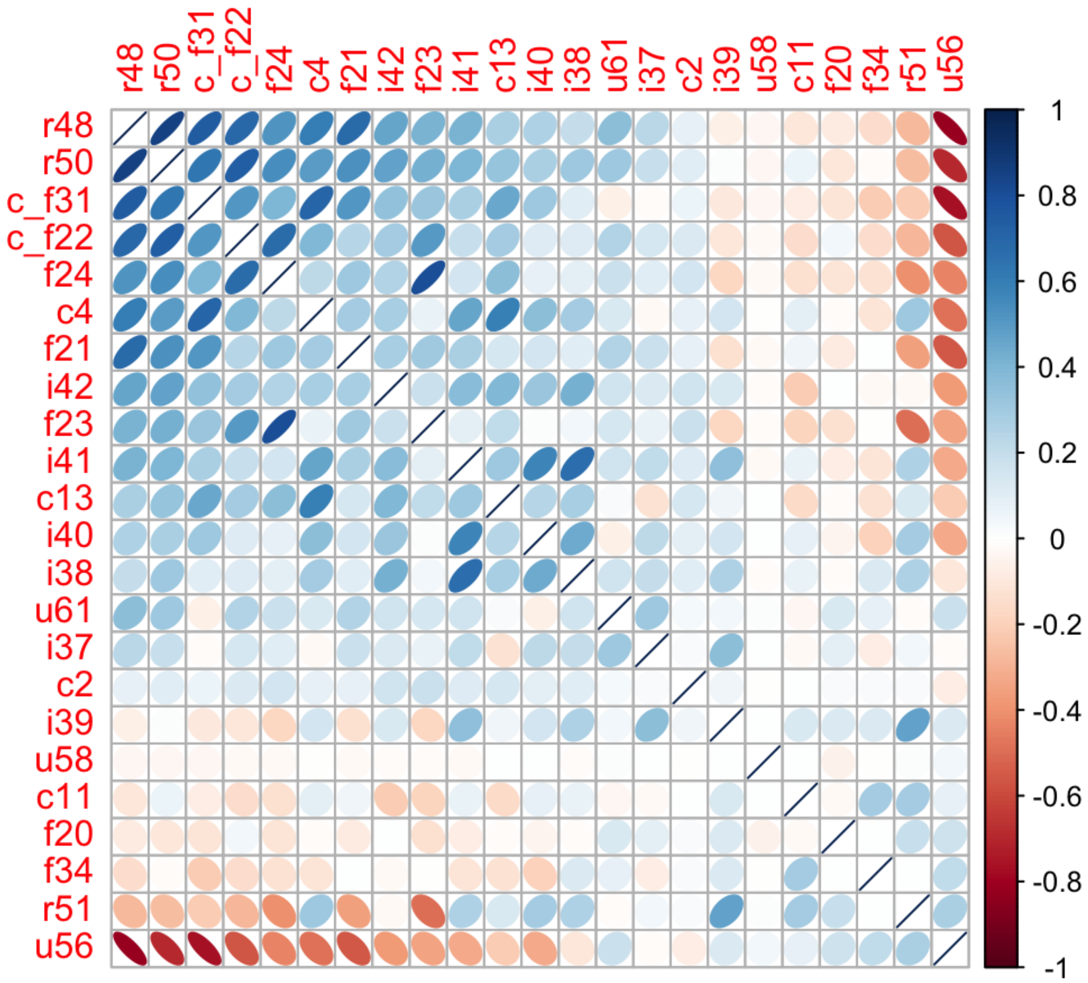


Figure 4.3: Correlations between the individual indicators. The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

Horn's parallel analysis revealed that six components must be extracted, as shown in Figure 4.4. The blue symbols × in the figure represent the eigenvalues for the observed correlation matrix, while the red dotted line indicates the mean of the corresponding eigenvalues for the simulated data.

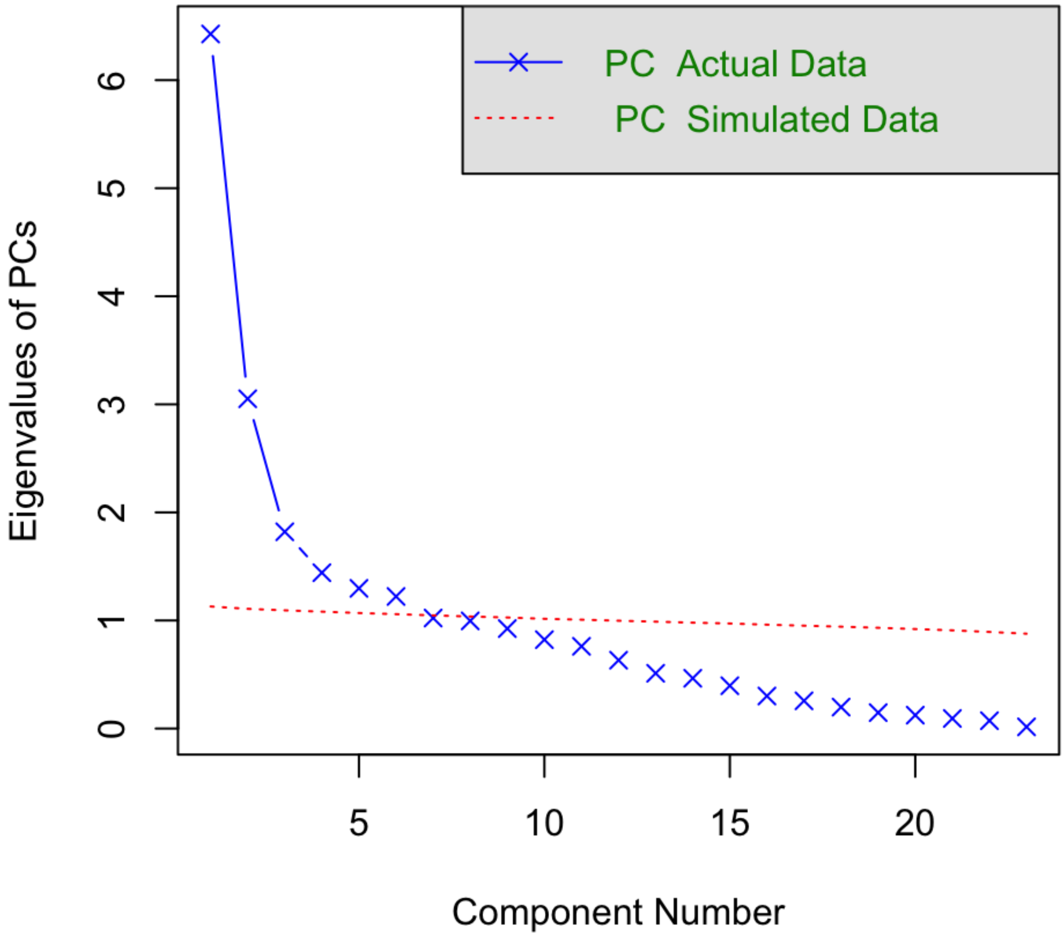


Figure 4.4: Results of the Horn's parallel analysis.

Table 4.10 shows the results of the principal components analysis. The loadings of the six rotated principal components are presented, with communalities of the individual indicators (h^2) and the variance explained by each principal component. As can be seen in the table, four individual indicators have communalities below 0.5. The individual indicator u58 has an extremely low communality of 0.013 and does not load on any of the six retained components. This is probably a consequence of the fact that only seven datasets have value of 1, and all other datasets have value of 0 (see Table 4.8). The other three indicators (c2, i42, i39) with low communalities load with high loadings on some principal component. Two individual indicators (c4, c13) have loadings above 0.4 on two components. The six retained principal components explain 66.3% variance. For this model, a fit based upon off diagonal values yielded a value of 0.95, which is larger than 0.9 and thus indicates a good fit. The root-mean-square residual reached a value of 0.06, which is smaller than 0.08, also indicating a good fit.

In order to better visualise the structure of the principal components, they are also presented in a network diagram (see Figure 4.5). In the figure, the individual indicators are represented by

Table 4.10: Loadings of the first six principal components ($RC1$ to $RC6$) after 'cluster' rotation with their variances and the individual indicators' communalities (h^2). The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

II	$RC1$	$RC2$	$RC5$	$RC3$	$RC4$	$RC6$	h^2
r48	0.967	-0.009	-0.023	-0.153	-0.002	0.175	0.946
c_f31	0.939	-0.059	-0.107	0.246	-0.062	-0.072	0.839
u56	-0.874	0.026	0.111	0.002	0.082	0.249	0.852
c4	0.817	0.234	-0.179	0.438	0.095	0.186	0.853
r50	0.813	0.074	0.185	-0.055	0.168	0.124	0.835
f21	0.776	-0.093	-0.096	-0.305	0.199	-0.001	0.630
c_f22	0.594	-0.132	0.350	0.111	-0.104	0.294	0.698
i38	-0.136	0.821	0.324	-0.027	0.144	-0.130	0.677
i41	0.173	0.799	0.052	-0.116	0.005	-0.144	0.731
i40	0.125	0.730	-0.115	-0.102	-0.171	-0.314	0.643
i39	-0.167	0.619	-0.143	-0.186	0.076	0.193	0.497
r51	-0.141	0.570	-0.397	0.212	0.083	0.186	0.800
i42	0.133	0.461	0.352	0.057	-0.144	-0.004	0.463
f23	0.037	-0.111	0.872	0.110	-0.010	-0.087	0.750
f24	0.213	-0.079	0.782	0.177	-0.065	0.016	0.760
c2	-0.236	0.235	0.603	0.298	0.117	-0.067	0.266
c13	0.199	0.280	0.433	0.705	-0.106	0.062	0.729
i37	-0.007	0.384	-0.163	-0.788	-0.210	0.292	0.706
f34	-0.106	-0.052	0.330	0.110	0.824	0.043	0.693
c11	0.303	0.046	-0.276	0.021	0.789	-0.060	0.674
f20	0.087	-0.107	-0.243	0.032	-0.154	0.735	0.544
u61	0.197	0.002	0.138	-0.320	0.124	0.690	0.662
u58	-0.120	0.057	0.063	-0.013	-0.014	-0.089	0.013
SS loadings	5.142	3.123	2.551	1.472	1.523	1.449	
Proportion Var	0.224	0.136	0.111	0.064	0.066	0.063	
Cumulative Var	0.224	0.359	0.470	0.534	0.601	0.663	

square nodes, the principal components by circular nodes, and the loadings by edges connecting them. Only loadings above 0.4 in absolute value are shown. The width of the edges is proportional to the absolute value of the corresponding loading. Green edge colour indicates a positive loading, while purple indicates a negative loading.

Since the first three principal components explain a significant portion of the variance in the data (47%), a separate biplot is created for each pair of principal components (see Figure 4.6). The projected observations (i.e., the open datasets) are shown with blue triangles (the European OGD portal) and red circles (the Australian OGD portal), while the projected individual indicators are shown as vectors. Triangles and circles that are close to each other represent observations

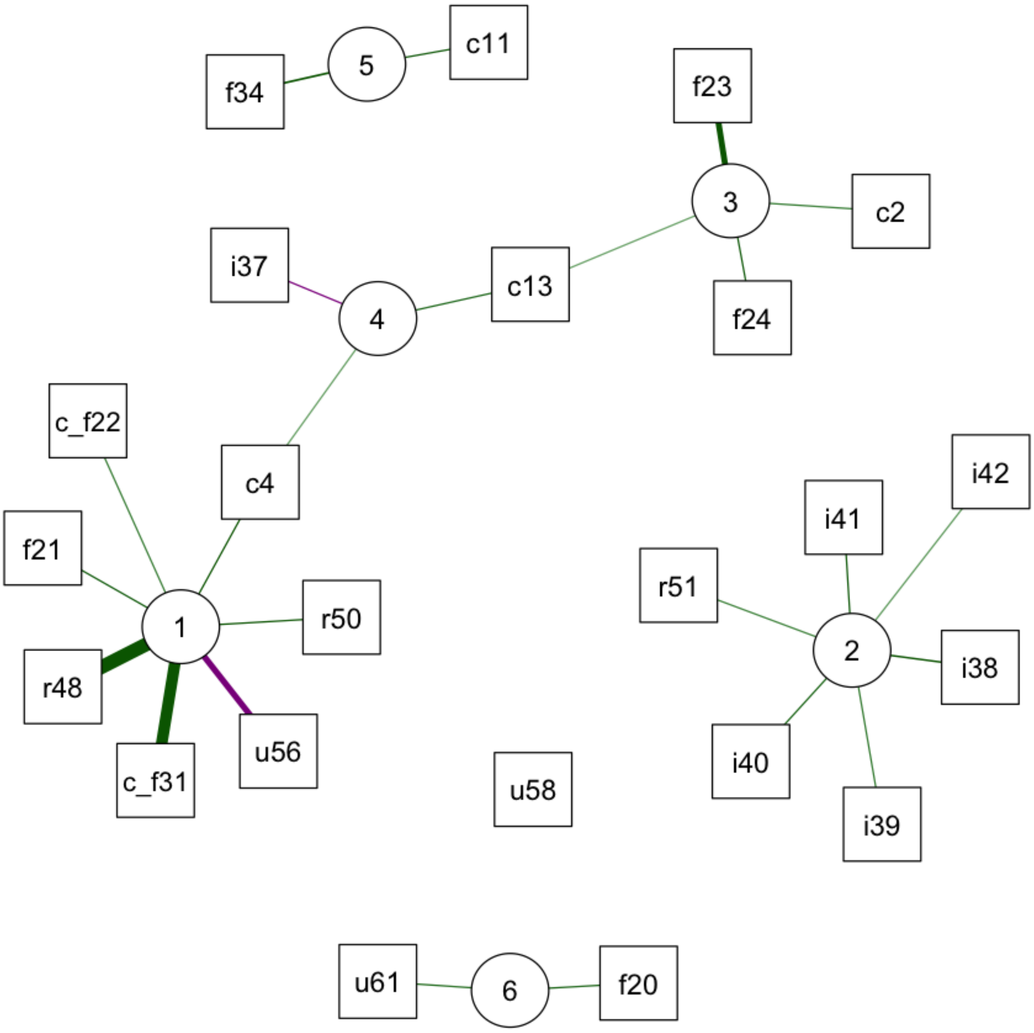


Figure 4.5: Loadings of individual indicators on principal components. The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

with similar values of the principal components. These biplots reveal a striking character of the first principal component (RC1), which almost perfectly distinguishes the two portals from each other. There is also a difference between the two portals in the third component (RC5), but it is not as striking.

As explained in subchapter 3.2.3 (page 31), no normalisation should be carried out when developing this composite indicator.

The weights of the dimensions (related to the properties of the open dataset) and the weights of the individual indicators are determined based on the analytic hierarchy process, as explained in subchapter 3.2.3 (page 31). Out of nine experts who have been invited, insights into the

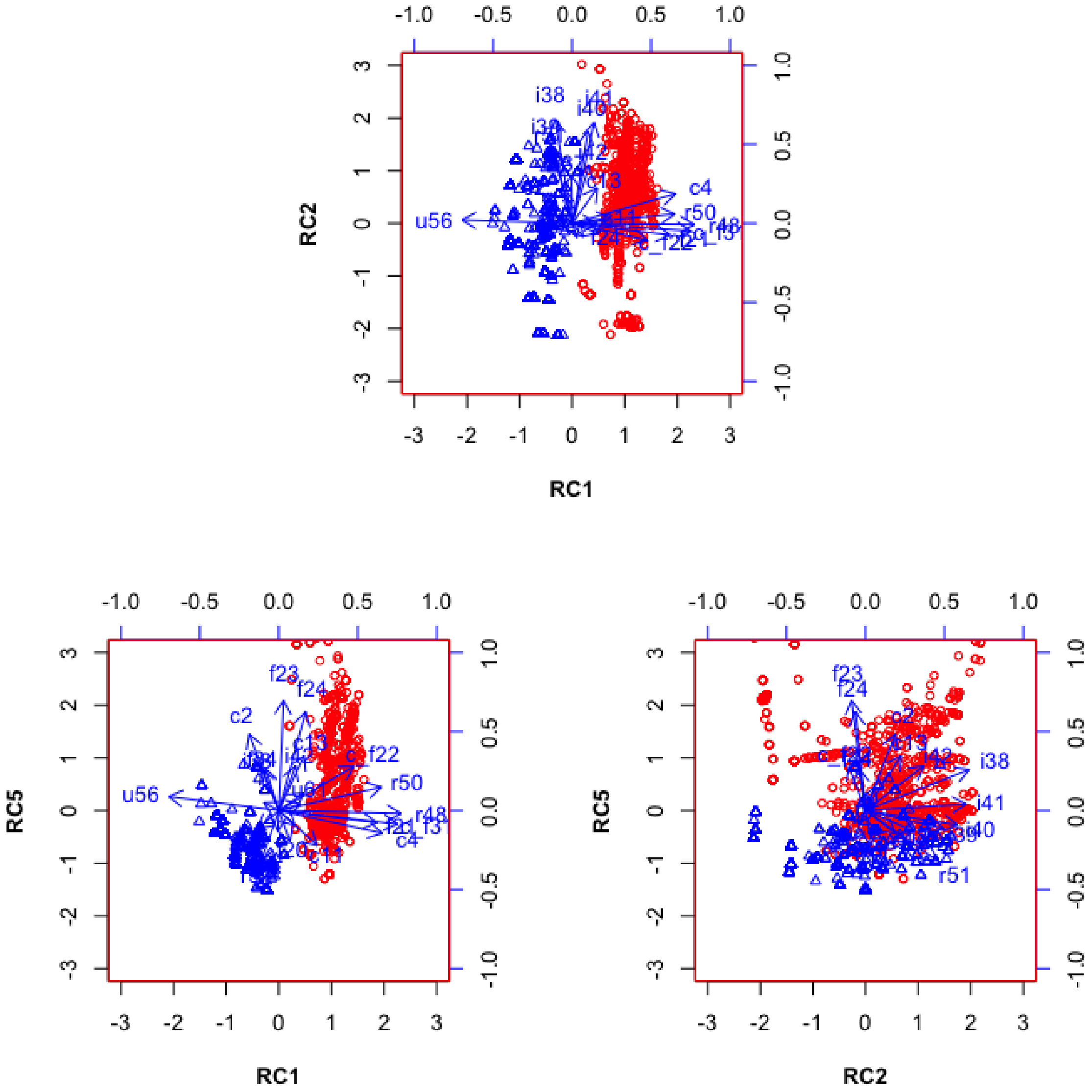


Figure 4.6: Pairwise biplots of the rotated first three principal components - PC1, PC2, and PC5. The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

relative importance of five dimensions and 32 individual indicators are provided by five of them (see Appendix C). The consistency of the experts' judgements regarding the relative importance of the dimensions and individual indicators is assessed by calculating the consistency ratios. Subsequently, the priorities/weights of the elements within the theoretical framework are calculated.

The priorities of the dimensions are calculated separately for each individual expert on the basis of their input, i.e., pairwise comparisons (see Table 4.11). The largest deviation in the

Table 4.11: Priorities and consistency ratios for dimensions (Legend: F - Findability, R - Retrievability, I - Interoperability, U - Reusability, C - Contextuality, CR - consistency ratio).

Expert	F	R	I	U	C	CR
expert1	0.07	0.13	0.26	0.22	0.32	0.01
expert2	0.41	0.20	0.13	0.06	0.20	0.02
expert3	0.36	0.10	0.13	0.08	0.33	0.07
expert4	0.39	0.39	0.09	0.09	0.04	0.02
expert5	0.37	0.30	0.14	0.12	0.07	0.04

priorities among different experts regarding the dimensions can be observed in the dimensions of findability, retrievability, and contextuality. For all experts, findability is the most important dimension, with the exception of expert 1, for whom it is the least important, both in comparison to the other experts and to the other dimensions. It is worth noting that for expert 4, findability and retrievability are equally important. Furthermore, contextuality is very important for experts 1, 2 and 3, while it is almost negligible for experts 4 and 5. As the consistency ratios of the pairwise comparison of the dimensions are less than 0.1 for each individual expert, the judgements are considered acceptable and no corrective measures are required. The pairwise comparisons of the dimensions of all five experts are aggregated using the geometric mean. The aggregated judgement matrix for the dimensions is presented in Table 4.12. The matrix is only listed for the sake of completeness, but the aggregated priorities as shown in Table 4.13 are relevant for the further development of the composite indicator. The consistency ratio of the aggregated judgements for the dimensions is 0.006, which is less than 0.1 and thus the aggregated judgements are considered acceptable. The priorities based on the aggregated judgements for the dimensions are listed in Table 4.13. These aggregated priorities represent the weightings assigned to the individual dimensions and show the relative importance of each dimension compared to others. Taking into account the differences in the priorities of the dimensions for the different experts (as shown in Table 4.11), the aggregated priorities listed in Table 4.13 are considered to be the consensus among the experts. Findability proved to be the most important dimension with an aggregate priority of 0.31, while reusability is the least important at 0.12.

Table 4.12: Aggregated judgement matrix for the dimensions (Legend: F - Findability, R - Retrievability, I - Interoperability, U - Reusability, C - Contextuality).

	F	R	I	U	C
F	1.00	1.43	2.02	2.32	1.95
R	0.70	1.00	1.72	2.09	1.16
I	0.49	0.58	1.00	1.78	1.00
U	0.43	0.48	0.56	1.00	0.79
C	0.51	0.86	1.00	1.27	1.00

The relative priorities of the individual indicators within the dimensions for each individual expert are shown in the following tables (for the individual indicators of findability – Table 4.14,

Table 4.13: Aggregated priorities for dimensions (Legend: F - Findability, R - Retrievability, I - Interoperability, U - Reusability, C - Contextuality).

F	R	I	U	C
0.31	0.23	0.17	0.12	0.17

for the individual indicators of retrievability – Table 4.15, for the individual indicators of interoperability – Table 4.16, for the individual indicators of reusability – Table 4.17, for the individual indicators of contextuality – Table 4.18) together with the consistency ratios. As the consistency ratios of the pairwise comparison of each individual indicator within the dimensions are less than 0.1 for each individual expert, the judgements are considered acceptable, and no corrective measures are required. The greatest difference in the priorities among different experts regarding the dimension of findability can be observed in the individual indicators f20 and f34. While f20 is the least important for expert 1 with a weighting of 0.06, it is the most important for expert 5 with a weighting of 0.38. Conversely, f34 is most important for expert 1 compared to the other experts (weight 0.43), while it is least important for expert 5 (weight 0.06). The greatest difference in the priorities among different experts regarding the dimension of retrievability can be observed in the individual indicator r51. The indicator r51 proved to be very important or the most important individual indicator for all experts compared to the other indicators within the dimension. The greatest difference in the priorities among different experts with regard to the dimension of interoperability can be observed in the individual indicator i41. In particular, experts 2 and 3 consider this individual indicator to be the most important, expert 4

Table 4.14: Priorities and consistency ratios for individual indicators of findability (Legend: CR - consistency ratio). The definitions of abbreviations for individual indicators are provided in Table 4.1, page 54.

Expert	f20	f21	f23	f24	f33	f34	CR
expert1	0.06	0.03	0.06	0.13	0.28	0.43	0.01
expert2	0.11	0.05	0.31	0.31	0.07	0.14	0.02
expert3	0.11	0.19	0.07	0.06	0.23	0.35	0.09
expert4	0.09	0.14	0.05	0.05	0.33	0.33	0.01
expert5	0.38	0.27	0.11	0.11	0.07	0.06	0.08

Table 4.15: Priorities and consistency ratios for individual indicators of retrievability (Legend: CR - consistency ratio). The definitions of abbreviations for individual indicators are provided in Table 4.2, page 56.

Expert	r48	r49	r50	r51	CR
expert1	0.22	0.04	0.08	0.66	0.05
expert2	0.06	0.12	0.26	0.57	0.03
expert3	0.48	0.14	0.09	0.29	0.06
expert4	0.17	0.33	0.33	0.17	0.00
expert5	0.42	0.29	0.17	0.12	0.04

Table 4.16: Priorities and consistency ratios for individual indicators of interoperability (Legend: CR - consistency ratio). The definitions of abbreviations for individual indicators are provided in Table 4.3, page 57.

Expert	i37	i38	i39	i40	i41	i42	CR
expert1	0.12	0.34	0.34	0.06	0.03	0.12	0.00
expert2	0.03	0.06	0.13	0.26	0.47	0.04	0.04
expert3	0.06	0.19	0.10	0.19	0.35	0.11	0.08
expert4	0.13	0.25	0.25	0.12	0.12	0.12	0.00
expert5	0.25	0.25	0.22	0.13	0.09	0.06	0.06

rates it as equally important compared to the other experts, while expert 1 considers it to be unimportant both compared to the other experts and compared to other indicators within the dimension. The greatest difference in the priorities among different experts with regard to the dimension of reusability can be observed in the individual indicators u61 and u_c3. Expert 1 and expert 2 in particular consider these individual indicators to be the most important. The largest difference in the priorities among different experts regarding the dimension of contextuality can be observed in the individual indicator c_f31. In particular, three experts (expert 1, expert 2,

Table 4.17: Priorities and consistency ratios for individual indicators of reusability (Legend: CR - consistency ratio). The definitions of abbreviations for individual indicators are provided in Table 4.4 and Table 4.5, pages 58-59. Indicators whose abbreviation begins with a letter followed by an underscore are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore.

Expert	u54	u55	u56	u58	u59	u61	u_c3	CR
expert1	0.10	0.05	0.10	0.10	0.10	0.37	0.19	0.00
expert2	0.03	0.04	0.07	0.10	0.16	0.24	0.35	0.02
expert3	0.17	0.15	0.19	0.20	0.07	0.08	0.13	0.09
expert4	0.12	0.20	0.12	0.20	0.12	0.12	0.10	0.04
expert5	0.20	0.18	0.21	0.16	0.12	0.07	0.06	0.09

Table 4.18: Priorities and consistency ratios for individual indicators of contextuality (Legend: CR - consistency ratio). The definitions of abbreviations for individual indicators are provided in Table 4.1 and Table 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore.

Expert	c2	c4	c9	c11	c12	c13	c18	c_f22	c_f31	CR
expert1	0.08	0.03	0.02	0.08	0.08	0.08	0.15	0.08	0.40	0.02
expert2	0.02	0.03	0.06	0.11	0.22	0.03	0.07	0.15	0.31	0.06
expert3	0.04	0.08	0.08	0.07	0.05	0.07	0.12	0.21	0.28	0.10
expert4	0.12	0.06	0.10	0.08	0.12	0.09	0.12	0.12	0.18	0.04
expert5	0.12	0.16	0.13	0.11	0.12	0.09	0.11	0.08	0.08	0.08

and expert 3) consider this indicator to be the most important. Expert 4 considers it moderately important (weight 0.18), while expert 5 considers it unimportant compared to the other experts (weight 0.08). Just as the pairwise comparisons of the dimensions of the individual experts are aggregated, the individual pairwise comparisons of the indicators are also aggregated separately for each dimension. This is done using the geometric mean. A separate aggregated judgement matrix is therefore created for the individual indicators of each dimension (the individual indicators of findability – Table 4.19, the individual indicators of retrievability – Table 4.20, the individual indicators of interoperability – Table 4.21, the individual indicators of reusability – Table 4.22, the individual indicators of contextuality – Table 4.23). The aggregated judgement matrices are only an intermediate result that is used to calculate the aggregated priorities of the individual indicators, but are listed here for the sake of completeness. The consistency ratio of the aggregated judgements for the individual indicators of findability is 0.018, for the individual indicators of retrievability is 0.001, for the individual indicators of interoperability 0.006, for the individual indicators of reusability 0.009, and for the individual indicators of contextuality 0.013.

Table 4.19: Aggregated judgement matrix for individual indicators of findability. The definitions of abbreviations for individual indicators are provided in Table 4.1, page 54.

	f20	f21	f23	f24	f33	f34
f20	1.00	1.08	1.58	1.38	0.72	0.39
f21	0.92	1.00	1.29	1.19	0.45	0.42
f23	0.63	0.77	1.00	0.87	0.75	0.61
f24	0.72	0.84	1.15	1.00	0.83	0.68
f33	1.38	2.22	1.33	1.20	1.00	0.76
f34	2.54	2.41	1.63	1.46	1.32	1.00

Table 4.20: Aggregated judgement matrix for individual indicators of retrievability. The definitions of abbreviations for individual indicators are provided in Table 4.2, page 56.

	r48	r49	r50	r51
r48	1.00	1.52	1.22	0.70
r49	0.66	1.00	0.92	0.54
r50	0.82	1.08	1.00	0.53
r51	1.42	1.86	1.89	1.00

Then, the aggregated priorities, which are the priorities calculated from the aggregated judgement matrices, are determined for each individual indicator within the dimensions of findability (Table 4.24), retrievability (Table 4.25), interoperability (Table 4.26), reusability (Table 4.27), and contextuality (Table 4.28). These aggregated priorities are seen as a consensus among the experts, as their opinions on the relative importance of the individual indicators differ (as shown in Table 4.14, Table 4.15, Table 4.16,Table 4.17, Table 4.18). Additionally, they offer valuable insights into the relative importance of each indicator within its dimension, facilitating

Table 4.21: Aggregated judgement matrix for individual indicators of interoperability. The definitions of abbreviations for individual indicators are provided in Table 4.3, page 57.

	i37	i38	i39	i40	i41	i42
i37	1.00	0.45	0.51	0.62	0.84	1.08
i38	2.22	1.00	1.06	1.70	1.01	2.17
i39	1.97	0.94	1.00	1.43	1.40	2.27
i40	1.60	0.59	0.70	1.00	0.92	1.89
i41	1.19	0.99	0.71	1.08	1.00	1.52
i42	0.92	0.46	0.44	0.53	0.66	1.00

Table 4.22: Aggregated judgement matrix for individual indicators of reusability. The definitions of abbreviations for individual indicators are provided in Table 4.4 and Table 4.5, pages 58-59. Indicators whose abbreviation begins with a letter followed by an underscore are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore.

	u54	u55	u56	u58	u59	u61	u_c3
u54	1.00	1.15	0.70	0.66	1.04	0.70	0.89
u55	0.87	1.00	1.00	0.70	1.08	0.86	0.53
u56	1.43	1.00	1.00	1.00	1.15	0.72	0.96
u58	1.52	1.43	1.00	1.00	1.55	1.00	0.94
u59	0.96	0.92	0.87	0.64	1.00	1.00	0.92
u61	1.43	1.16	1.40	1.00	1.00	1.00	1.25
u_c3	1.12	1.89	1.05	1.06	1.08	0.80	1.00

Table 4.23: Aggregated judgement matrix for individual indicators of contextuality. The definitions of abbreviations for individual indicators are provided in Table 4.1 and Table 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore.

	c2	c4	c9	c11	c12	c13	c18	c_f22	c_f31
c2	1.00	1.00	1.23	0.68	0.56	0.92	0.57	0.56	0.29
c4	1.00	1.00	1.32	0.72	0.47	1.22	0.39	0.41	0.27
c9	0.81	0.76	1.00	0.78	0.65	1.36	0.45	0.39	0.36
c11	1.48	1.38	1.28	1.00	1.06	1.64	0.76	0.56	0.35
c12	1.78	2.11	1.54	0.94	1.00	1.22	1.06	1.00	0.53
c13	1.08	0.82	0.74	0.61	0.82	1.00	0.66	0.76	0.43
c18	1.74	2.55	2.22	1.32	0.94	1.52	1.00	0.90	0.44
c_f22	1.78	2.46	2.57	1.78	1.00	1.32	1.11	1.00	0.49
c_f31	3.44	3.76	2.77	2.83	1.89	2.35	2.27	2.05	1.00

the further development of composite indicators. The most important individual indicator of the findability dimension is f34 with an aggregated priority of 0.26, while the least important is f23 with 0.12. The most important individual indicator of the retrievability dimension is r51 with an aggregated priority of 0.36, while the least important is r49 with 0.18. In the dimension of interoperability, the highest aggregated priority of 0.23 is assigned to two individual indicators, i38 and i39. On the other hand, i42, with an aggregated priority of 0.1, proved to be the least important indicator of the interoperability dimension. The most important individual indicator of the reusability dimension is u58 with an aggregated priority of 0.17, whereas two indicators, u54 and u55, received the lowest aggregated priority value of 0.12. One indicator, c_f31, of the contextuality dimension achieved the highest aggregated priority value of 0.24, while three indicators, c2, c4 and c9, received the lowest value of 0.07.

Table 4.24: Aggregated priorities for individual indicators of findability. The definitions of abbreviations for individual indicators are provided in Table 4.1, page 54.

f20	f21	f23	f24	f33	f34
0.15	0.13	0.12	0.14	0.20	0.26

Table 4.25: Aggregated priorities for IIs of individual indicators of retrievability. The definitions of abbreviations for individual indicators are provided in Table 4.2, page 56.

r48	r49	r50	r51
0.26	0.18	0.20	0.36

Table 4.26: Aggregated priorities for individual indicators of interoperability. The definitions of abbreviations for individual indicators are provided in Table 4.3, page 57.

i37	i38	i39	i40	i41	i42
0.11	0.23	0.23	0.16	0.17	0.10

Table 4.27: Aggregated priorities for individual indicators of reusability. The definitions of abbreviations for individual indicators are provided in Table 4.4 and Table 4.5, pages 58-59. Indicators whose abbreviation begins with a letter followed by an underscore are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore.

u54	u55	u56	u58	u59	u61	u_c3
0.12	0.12	0.14	0.17	0.13	0.16	0.16

Table 4.28: Aggregated priorities for individual indicators of contextuality. The definitions of abbreviations for individual indicators are provided in Table 4.1 and Table 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore.

c2	c4	c9	c11	c12	c13	c18	c_f22	c_f31
0.07	0.07	0.07	0.10	0.12	0.08	0.13	0.14	0.24

As already described in subchapter 3.2.3 (page 31), linear aggregation has been proven to be a suitable method for aggregating elements of the theoretical framework of metadata quality for open datasets. As an example, the aggregated values/scores of the dimensions and the values/scores of the composite indicator for some open datasets are provided in Table 4.29. In addition, the results of the linear aggregation across relevant individual indicators to dimensions and these to a composite indicator for all downloaded open datasets are stored on the Harvard Dataverse (Šlibar, 2024a).

Table 4.29: An example of the aggregated scores for some open datasets (Legend: C - Contextuality, F - Findability, I - Interoperability, R - Retrievability, U - Reusability, CI - Composite indicator).

Dataset	**C**	**F**	**I**	**R**	**U**	**CI**
645	0.2149	0.6127	0.7673	0.3606	0.0000	0.4406
679	0.5967	0.6127	0.7303	0.3606	0.3087	0.5349
605	0.5967	0.3488	0.3387	0.3606	0.1438	0.3676
627	0.3594	0.6127	0.8571	0.3606	0.1438	0.4966
668	0.3594	0.6127	0.8994	0.3606	0.0000	0.4868
619	0.5198	0.6127	0.5011	0.3606	0.3087	0.4839

4.2.2 Demonstrate artefact

The values of the lower- and higher-level aggregates of the developed composite indicator (dimensions and the composite indicator), along with the individual indicator values, are illustrated using an open dataset as an example (see Table 4.30).

Table 4.30: An example of computing the composite indicator for one open dataset (Legend: II - Individual indicator, Dim - Dimension, CI - Composite indicator). The definitions of abbreviations for individual indicators are provided in Tables 4.1, 4.2, 4.3, 4.4, 4.5, pages 54-59. Indicators whose abbreviation begins with a letter followed by an underscore (e.g., c_f22) are the ones that changed the dimension in the final theoretical framework. In the previously mentioned tables, they can be identified by the last part of the abbreviation, i.e., the part after the underscore (e.g., f22).

Metadata	II		Dim		CI
canberra; kiss and ride; light rail; public transport; transport	f20	1	F	0.74	0.73
Transport	f21	1			
NA	f23	0			
NA	f24	0			
https://data.gov.au/dataset/ds-act-https://www.data.act.gov.au/api/views/2q44-5we7	f33	1			
https://data.gov.au/dataset/ds-act-https://www.data.act.gov.au/api/views/2q44-5we7	f34	1			

Metadata	II		Dim		CI
https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json	r48	1	R	1	0.73
https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json	r49	1			
https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json; https://www.data.act.gov.au/data.json	r50	1			
https://www.data.act.gov.au/api/views/2q44-5we7/rows.csv?accessType=DOWNLOAD; https://www.data.act.gov.au/api/views/2q44-5we7/rows.rdf?accessType=DOWNLOAD; https://www.data.act.gov.au/api/views/2q44-5we7/rows.json?accessType=DOWNLOAD; https://www.data.act.gov.au/api/views/2q44-5we7/rows.xml?accessType=DOWNLOAD	r51	1			
CSV;RDF;JSON;XML	i37	1	I	1	
CSV;RDF;JSON;XML	i38	1			
CSV;RDF;JSON;XML	i39	1			
CSV;RDF;JSON;XML	i40	1			
CSV;RDF;JSON;XML	i41	1			
text/csv; application/rdf+xml; application/json; application/xml	i42	1			
NA	u54	0	U	0.16	
NA	u55	0			
NA	u56	0			
NA	u58	0			
NA	u59	0			

Metadata	II		Dim		CI
http://creativecommons.org/licenses/by-sa/4.0/legalcode; http://creativecommons.org/licenses/by-sa/4.0/legalcode; http://creativecommons.org/licenses/by-sa/4.0/legalcode; http://creativecommons.org/licenses/by-sa/4.0/legalcode	u61	1	U	0.16	0.73
NA	u_c3	0			
NA AND NA	c2	0	C	0.45	
NA	c4	0			
NA	c9	0			
2021-11-25T13:00Z	c11	1			
2021-11-25T13:00Z	c12	1			
NA	c13	0			
NA	c18	0			
NA	c_f22	0			
Kiss And Ride Locations	c_f31	1			

The equations below explain how to calculate the scores for all aggregates (dimensions and the composite indicator) of the composite indicator. This involves adding the weights of the individual indicators multiplied by their scores and adding the weights of the dimensions multiplied by their scores. The dataset from Table 4.30 is used to illustrate this process.

$$\begin{aligned}
F &= 0.15 \cdot 1 + 0.13 \cdot 1 + 0.12 \cdot 0 + 0.14 \cdot 0 + 0.2 \cdot 1 + 0.26 \cdot 1 \\
R &= 0.26 \cdot 1 + 0.18 \cdot 1 + 0.2 \cdot 1 + 0.36 \cdot 1 \\
I &= 0.11 \cdot 1 + 0.23 \cdot 1 + 0.23 \cdot 1 + 0.16 \cdot 1 + 0.17 \cdot 1 + 0.1 \cdot 1 \\
U &= 0.12 \cdot 0 + 0.12 \cdot 0 + 0.14 \cdot 0 + 0.17 \cdot 0 + 0.13 \cdot 0 + 0.16 \cdot 1 + 0.16 \cdot 0 \\
C &= 0.07 \cdot 0 + 0.07 \cdot 0 + 0.07 \cdot 0 + 0.1 \cdot 1 + 0.12 \cdot 1 + 0.08 \cdot 0 + 0.13 \cdot 0 + 0.14 \cdot 0 + 0.24 \cdot 1 \\
CI &= 0.31 \cdot 0.74 + 0.23 \cdot 1 + 0.17 \cdot 1 + 0.12 \cdot 0.16 + 0.17 \cdot 0.45
\end{aligned}$$

The descriptive statistics for the aggregates (scores of the dimensions and the composite indicator) of the developed composite indicator of all observed open datasets retrieved from the Australian OGD portal are presented in Table 4.31 and from the European OGD portal in Table 4.32. The composite scores of the datasets retrieved from the Australian OGD portal range between 0.343 and 0.869. Since the mean value for the retrievability dimension is very high, namely 0.902, it can be said that the vast majority of datasets on the Australian OGD portal achieve a very high quality according to this dimension. This is also reflected in the first quartile value,

Table 4.31: Descriptive statistics of the aggregates of the composite indicator for the Australian OGD portal (Legend: Min. - Minimum, Q1 - First quartile, $\bar{x}$ - Mean, Q3 - Third quartile, Max. - Maximum, SD - Standard deviation, IQR - Interquartile range).

	Min.	*Q1*	*Median*	$\bar{x}$	*Q3*	*Max.*	*SD*	*IQR*
Contextuality	0.375	0.520	0.597	0.620	0.723	0.804	0.093	0.203
Findability	0.199	0.613	0.741	0.748	0.872	1.000	0.165	0.259
Interoperability	0.102	0.314	0.508	0.552	0.858	1.000	0.307	0.545
Retrievability	0.440	0.800	1.000	0.902	1.000	1.000	0.153	0.200
Reusability	0.165	0.165	0.165	0.165	0.165	0.165	0.000	0.000
Composite indicator	0.343	0.593	0.659	0.661	0.735	0.869	0.097	0.142

which is 0.8 for retrievability, meaning that more than 75% of Australian datasets achieve the best quality for retrievability. Moreover, the standard deviation of reusability is 0, which means that the reusability scores are the same for all datasets retrieved from the Australian OGD portal. On the other hand, interoperability has the highest standard deviation (0.307), which means that the Australian datasets vary the most in this dimension. Although interoperability is assigned a lower weight, i.e., aggregated priority, compared to the other dimensions (as shown in Table 4.13), this dimension contributes to the variability of the composite scores on the Australian OGD portal given the high standard deviation. The composite scores of the datasets retrieved from the European OGD portal range between 0.182 and 0.585. The first quartile value is 0.613 for the findability dimension, which means that more than 75% of the datasets on the European OGD portal achieve the best quality in this dimension. Since the first quartile, the third quartile, and the median for findability are the same (0.613), this means that at least 50% of the datasets on the European OGD portal have the same findability score. In addition, the standard deviation of findability is very low (0.07), indicating minimal variability between the European datasets. A similar pattern can be observed for retrievability - the first quartile, third quartile, and median value is 0.316 and the standard deviation is 0.051. While the first quartile, third quartile and median values for the reusability dimension are identical (0.309), the standard deviation is not as low as for the findability and retrievability dimensions, indicating greater variability in the European datasets in terms of reusability. Like the Australian datasets, the European datasets

Table 4.32: Descriptive statistics of the aggregates of the composite indicator for the European OGD portal (Legend: Min. - Minimum, Q1 - First quartile, $\bar{x}$ - Mean, Q3 - Third quartile, Max. - Maximum, SD - Standard deviation, IQR - Interquartile range).

	Min.	*Q1*	*Median*	$\bar{x}$	*Q3*	*Max.*	*SD*	*IQR*
Contextuality	0.215	0.215	0.215	0.306	0.452	0.597	0.147	0.237
Findability	0.349	0.613	0.613	0.597	0.613	0.872	0.070	0.000
Interoperability	0.000	0.309	0.339	0.379	0.422	0.899	0.227	0.113
Retrievability	0.000	0.361	0.361	0.353	0.361	0.361	0.051	0.000
Reusability	0.000	0.309	0.309	0.241	0.309	0.475	0.125	0.000
Composite indicator	0.182	0.406	0.406	0.413	0.436	0.585	0.052	0.031

also show the greatest variation in terms of interoperability (with a standard deviation of 0.227). Although interoperability is assigned a lower weight, i.e., aggregated priority, compared to the other dimensions (as shown in Table 4.13), this dimension contributes the most to the variability of the composite scores on the European OGD portal given the high standard deviation of this dimension. And although findability is assigned the highest weight / aggregated priority compared to the other dimensions (as in Table 4.13), this dimension contributes only moderately to the variability of the composite scores on the European OGD portal given the very low standard deviation of this dimension.

The distributions of the aggregates (dimension scores and composite indicator scores) for all observed open datasets retrieved from the Australian OGD portal and the European OGD portal of the developed composite indicator are visualised using boxplots (see Figure 4.7). In the middle of each boxplot there is a horizontal line that represents the median of the visualised scores of one aggregate (e.g., composite indicator, contextuality, findability) for the Australian and European open datasets. In addition, datasets that deviate from the others (i.e., outliers) and reach a different score in relation to a certain aggregate are represented by dots that appear lighter when such datasets are fewer and darker when their frequency increases. Figure 4.7 shows that Australian datasets achieve higher scores, or have better metadata quality, in all aggregates compared to European datasets, with the exception of the reusability dimension, where the European datasets are slightly better than the Australian datasets. This trend is consistent with the data presented in Table 4.31 and Table 4.32. Moreover, Figure 4.7 shows that greater variability occurs when the box representing the interquartile range is longer. This relationship is also confirmed by the data in Table 4.31 and Table 4.32, which also contain information about the interquartile range, the first quartile, and the third quartile.

The structure of the composite indicator for four individual open datasets is shown in Figure 4.8 as radial plots: one with a high composite score from the Australian OGD portal (top left), one with a low composite score from the Australian OGD portal (top right), one with a high composite score from the European OGD portal (bottom left) and one with a low composite score from the European OGD portal (bottom right). In these plots, the grey line shows the average dimension scores of all observed datasets on the portal, while the red line shows the dimension scores of the individual dataset. The plots show that an individual dataset can be above average in some dimensions and below average in others. The datasets therefore do not have to have a uniform metadata quality. They can be very good in one dimension and very poor in another, compared to other datasets in the portal sample. The Australian dataset with a high composite score (top left) has above average scores for interoperability and retrievability, average scores for findability and reusability, and a below average score for contextuality. The Australian dataset with a low composite score (top right) has below average scores for interoperability, retrievability, contextuality, and findability, while it has an average score for reusability. The European dataset with a high composite score (bottom left) has above average scores for findability, contextuality, and reusability, while it has average scores for

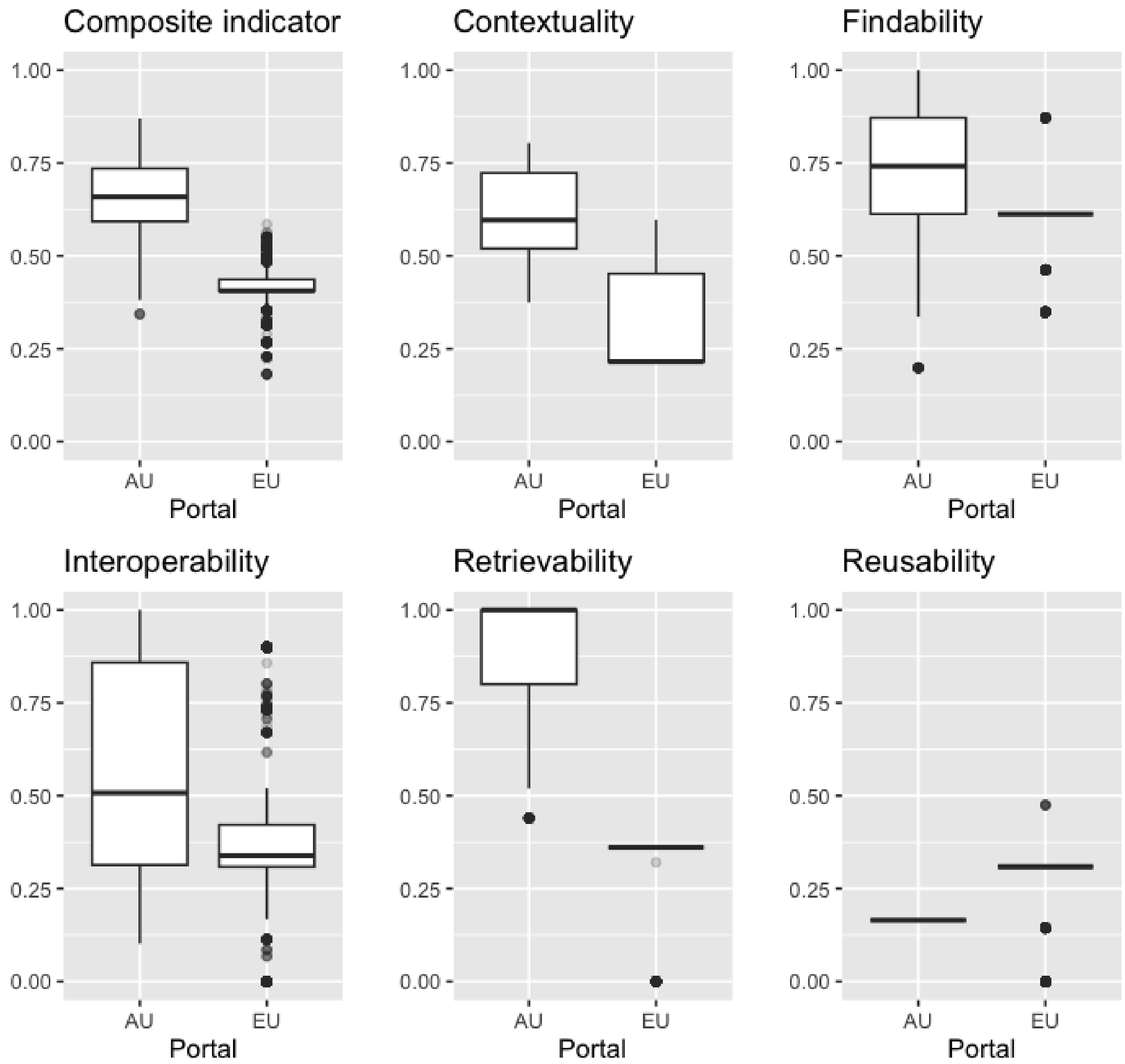


Figure 4.7: Distributions of the aggregates for all observed Australian (AU) and European (EU) open datasets.

interoperability and retrievability. The European dataset with a low composite score (bottom right) has below average scores for interoperability, retrievability, and reusability, an average score for findability, and an above average score for contextuality.

4.2.3 Evaluate artefact

The results of the 1st uncertainty analysis show how final composite values (scores and ranks) associated with each dataset will behave in different simulations due to changes in the weighting of the elements of the composite indicator and the use of different aggregation methods during the construction of the composite indicator (as described in subchapter 3.2.5, page 45).

The distributions of scores (see Figure 4.9) and ranks (see Figure 4.10) in the different simulations for every 10th dataset (from a subset of 679 datasets with unique scores of individual indicators) are represented by density curves in the violin plots. In the centre of each density

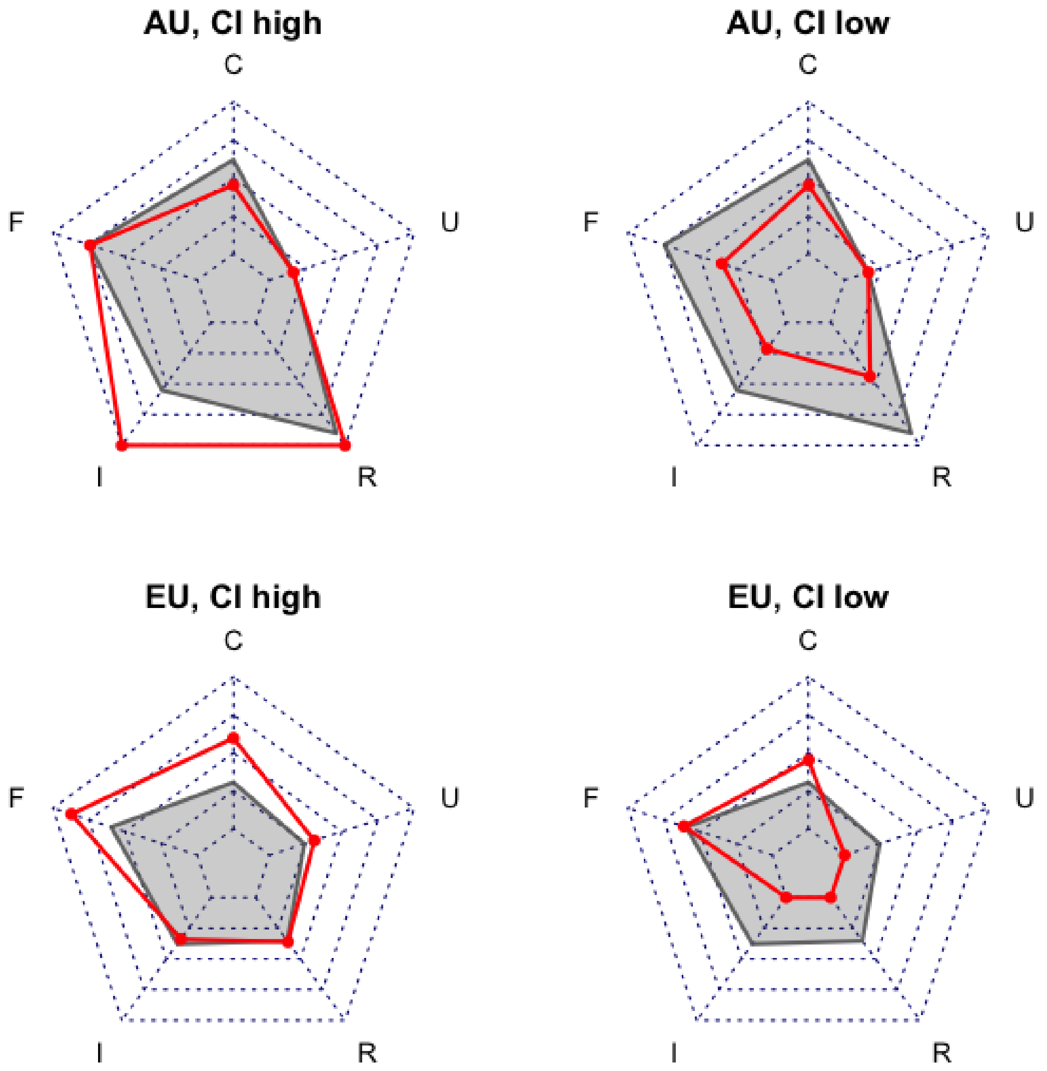


Figure 4.8: Structure of the composite indicator for individual open datasets (Legend: CI - Composite indicator, AU - Australian OGD portal, EU - European OGD portal).

curve is a small box representing the first and third quartiles, and a central red dot indicates the median. Each violin in Figure 4.9 is shaped like an hourglass. This shape is the result of the chosen aggregation method. Depending on which aggregation method is selected when constructing the composite indicator, the datasets in different simulations will achieve a lower or higher score for the composite indicator. When the weighted geometric mean (i.e., geometric aggregation) is chosen as the aggregation method, the dataset scores lower in all simulations (the bottom bulb of the hourglass), and when the weighted arithmetic mean (i.e., linear aggregation) is chosen, the dataset scores higher in all simulations (the top bulb of the hourglass). In addition, Figure 4.9 shows faster growth in scores (on the far left side), suggesting that the composite indicator developed can effectively distinguish between relatively similar datasets that have achieved a low score. The same is true for datasets that have achieved a high score (on the far-right side). The slower growth in scores (in the centre) suggests that the unperturbed composite

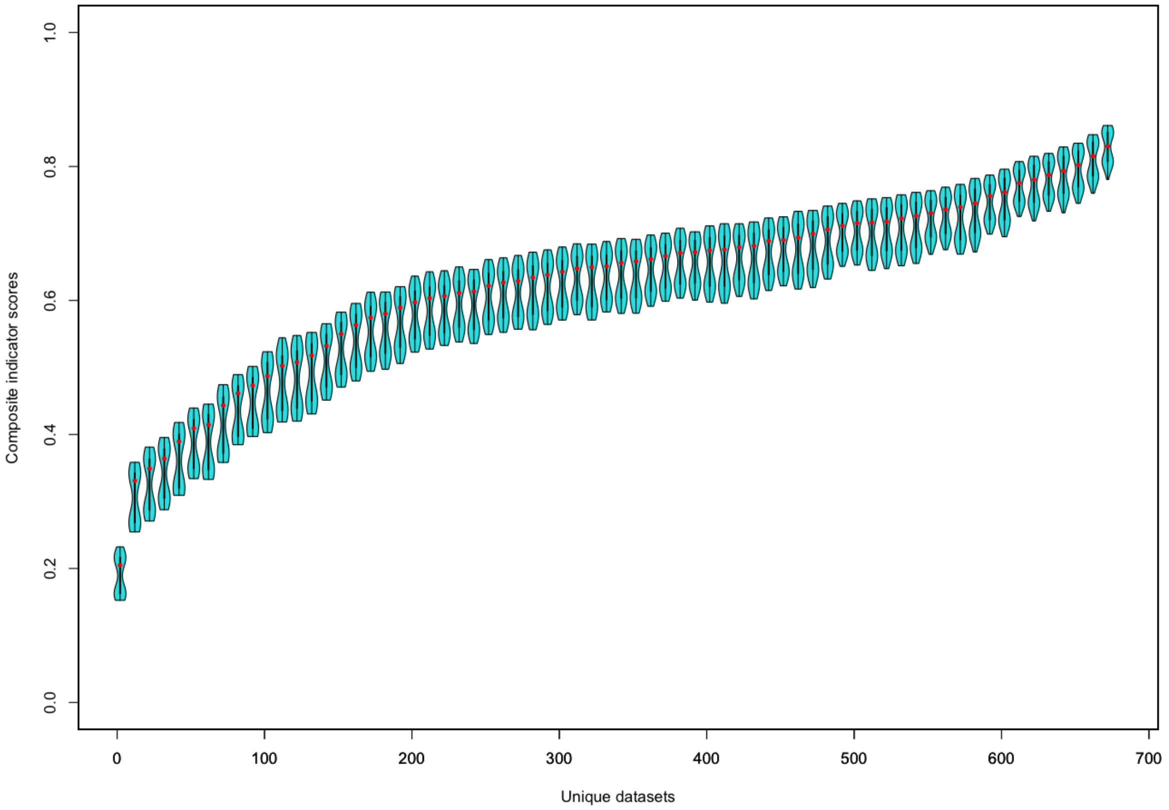


Figure 4.9: Distributions of the scores of the composite indicator assigned to the open datasets in different replications during the 1st uncertainty analysis. The datasets are ordered by the increasing score of the unperturbed composite indicator.

indicator is less effective at distinguishing between relatively similar datasets that have achieved a medium score. It can be observed that the violin plots in Figure 4.10 do not have the hourglass shape that can be seen in Figure 4.9. This is due to the fact that the ranking of the datasets in the different simulations is not strongly influenced by two input uncertainties (perturbation of the weights and selection of the aggregation method). In addition, it can be seen in Figure 4.10 that datasets that are mid-ranked according to the unperturbed composite indicator are more susceptible to change (as indicated by the slimmer violins and their elongated tails).

Sobol' sensitivity measures, specifically addressing two input uncertainties - the perturbation of weights (p_w) and the choice of the aggregation method (c_a) - are estimated. This estimation concentrates on the average absolute rank change between original/unperturbed and perturbed values. The estimates generated through the Monte Carlo method are presented in Table 4.33. Furthermore, the 90% confidence intervals for these measures are also calculated. The estimation of confidence intervals employs bootstrapping, underscoring the reliability of the total effect sensitivity indices' (S_{Ti}) estimates. On the other hand, the confidence intervals for the first-order sensitivity indices (S_i) are notably wider. However, given the number of runs conducted, robust conclusions can still be drawn. It can be observed that the first-order sensitivity index (S_i) for

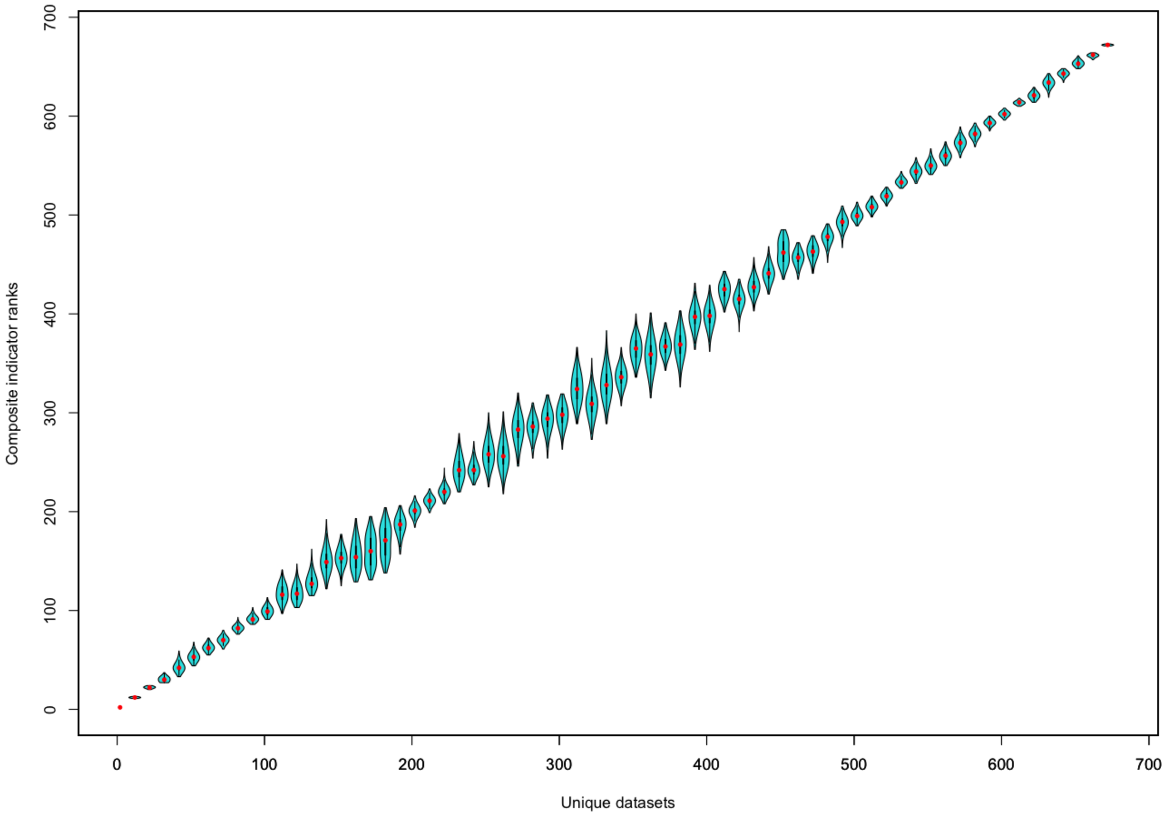


Figure 4.10: Distributions of the ranks of the composite indicator assigned to the open datasets in different replications during the 1st uncertainty analysis. The datasets are ordered by the increasing rank of the unperturbed composite indicator.

Table 4.33: Sobol' sensitivity measures and 90% confidence intervals estimated with the Monte Carlo method for all input assumptions. (Legend: S_i - First-order sensitivity index, S_{Ti} - Total effect sensitivity index, S_i_q5 - 5th percentile of the first-order sensitivity index, S_i_q95 - 95th percentile of the first-order sensitivity index, S_{Ti}_q5 - 5th percentile of the total effect sensitivity index, S_{Ti}_q95 - 95th percentile of total effect sensitivity index).

Input assumption	S_i	S_{Ti}	S_i_q5	S_i_q95	S_{Ti}_q5	S_{Ti}_q95
Perturbation of the weights (p_w)	0.2226	0.3858	0.1392	0.3094	0.3669	0.4054
Choice of aggregation method (c_a)	0.5391	0.6783	0.4241	0.6517	0.6550	0.7031

input uncertainty c_a is higher than for p_w. This indicates that the direct contribution of c_a to the variance of the output is larger than that of p_w. Essentially, fixing c_a at a particular value would, on average, lead to a greater reduction in the variance of output than fixing p_w. This sensitivity measure reflects the main effect of each input on the output of the model. The total effect sensitivity index (S_{Ti}) is also higher for c_a than for p_w. This indicates that the influence of c_a on the variance of the output is higher than the influence of p_w when both direct and indirect contributions (including interactions with other inputs) are taken into account. The total effect sensitivity index therefore captures the total effect of an input uncertainty on the variability of output, taking into account all possible ways in which the input influences the output. The sum of the first-order sensitivity indices is 0.76, while the sum of the total effect sensitivity indices is 1.06. As these two sums both differ from 1, there must be interactions between the input uncertainties in the model. Since both inputs p_w and c_a have total effect sensitivity indices that are greater than their first-order sensitivity indices, it can be concluded that they are involved in interactions. The difference $S_{Ti} - S_i$ of the input p_w is larger than that of the input c_a, which means that the input assumption p_w is more involved in the interactions. Although the value of S_{Ti} for the input p_w is greater than for the input c_a, both inputs are further analysed independently of each other.

The results of the 2nd uncertainty analysis show how final composite values associated with each dataset will behave in different simulations due to changes in the weighting of individual indicators resulting from the removal of a particular individual indicator (as described in subchapter 3.2.5, page 45). The distributions of scores (see Figure 4.11) and ranks (see Figure 4.12) in different simulations for every 10th dataset from a subset of 670 datasets with unique values of individual indicators are visualised using violin plots. Figure 4.11 shows that datasets with a low (those to the left) or medium (those in the middle) score according to the unperturbed

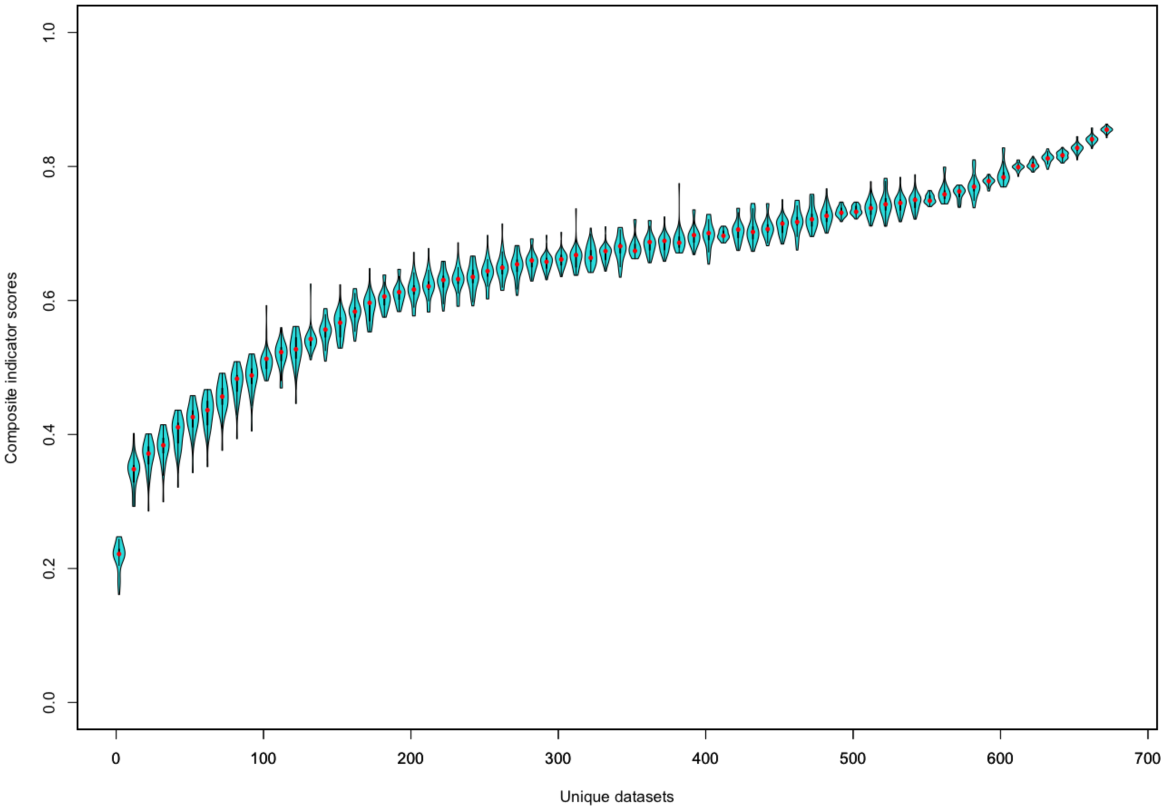


Figure 4.11: Distributions of the scores of the composite indicator assigned to the open datasets in different replications during the 2nd uncertainty analysis. The datasets are ordered by the increasing score of the unperturbed composite indicator.

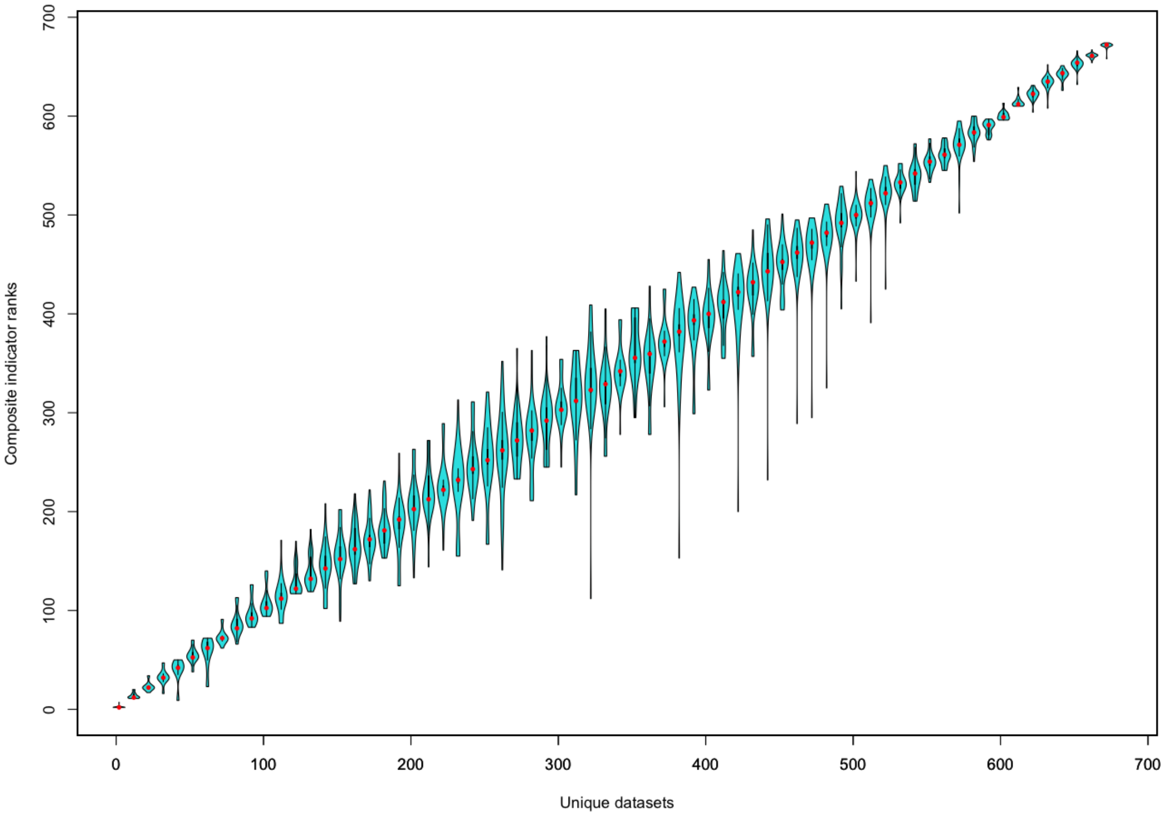


Figure 4.12: Distributions of the ranks of the composite indicator assigned to the open datasets in different replications during the 2nd uncertainty analysis. The datasets are ordered by the increasing rank of the unperturbed composite indicator.

composite indicator are affected by the removal of an individual indicator. In contrast, this change has no major impact on the datasets with the highest scores (those to the right). The greater sensitivity of medium quality datasets according to the unperturbed composite indicator is also evident in Figure 4.12. In the figure, the highest-ranked datasets (those on the left) and the lowest-ranked datasets (those on the right) show more consistent ranking across different replications, while the middle-ranked datasets (those in the centre) show greater variability (as indicated by the slimmer violins and their elongated tails).

The 3rd uncertainty analysis examined the robustness of the composite indicator in relation to the use of different aggregation methods in the construction of the composite indicator (as described in subchapter 3.2.5, page 45). Pearson's correlation coefficient between composite indicator scores determined that the use of two aggregation methods (weighted arithmetic mean and weighted geometric mean) is 0.99883 (95% confidence interval = 0.99876, 0.99889). In addition, Pearson's correlation coefficient was calculated for the ranks of the composite indicator, which were determined using two aggregation methods. It is 0.99958 (95% confidence interval = 0.99956, 0.99961). The calculated correlations are shown in Figure 4.13. The correlation coefficient between the composite indicator scores determined using the two aggregation

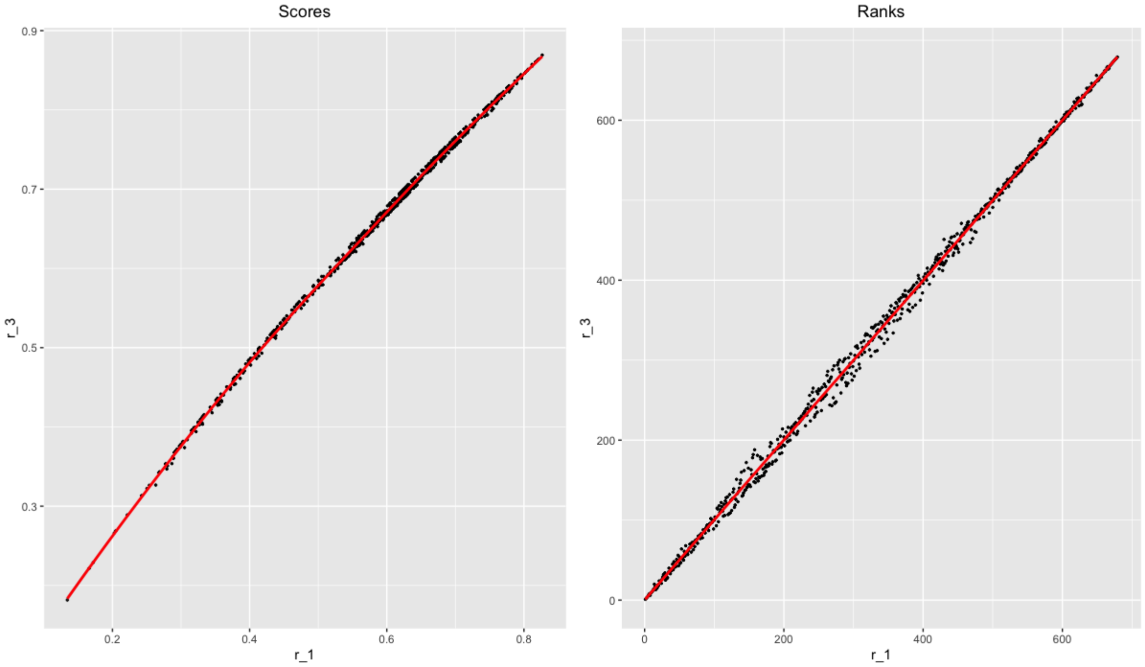


Figure 4.13: Scatterplots of the results of the geometric vs. linear aggregation method for composite indicator scores (left) and rank (right).

methods in different simulations is 0.99883 with a p-value of 0.00000 and is consistent with the hypothesis that the correlation coefficient between the composite indicator scores is larger than 0.99. The correlation coefficient between the ranks of the composite indicator, which was determined using the two aggregation methods is 0.99958 with a p-value of 0.00000 and is also consistent with the hypothesis that the correlation coefficient between the ranks of the composite indicator is larger than 0.99.

The 4th uncertainty analysis examined the robustness of the composite indicator in relation to changes in the weights of the elements at all levels of the composite indicator (as described in subchapter 3.2.5, page 45). For a total of 30 different perturbation weighting coefficients, 60,000 replications of the composite indicator values were generated. Therefore, for each selected perturbation coefficient, a total of 230,181 pairs of datasets were observed, and it was analysed whether the ranking order of each pair of datasets changed in the different replications. For the smallest observed weighting perturbation coefficient of 0.01, a total of 2.56% of dataset pairs changed their ranking order, for the coefficient of 0.05, this change occurred in 12.6% of pairs, and for the largest observed weighting perturbation coefficient of 0.95, the change occurred in 85.22% of pairs. In addition, the percentage of dataset pairs in which the ranking relationship changes due to an increase in the coefficient of the weighting perturbation is shown graphically (see Figure 4.14).

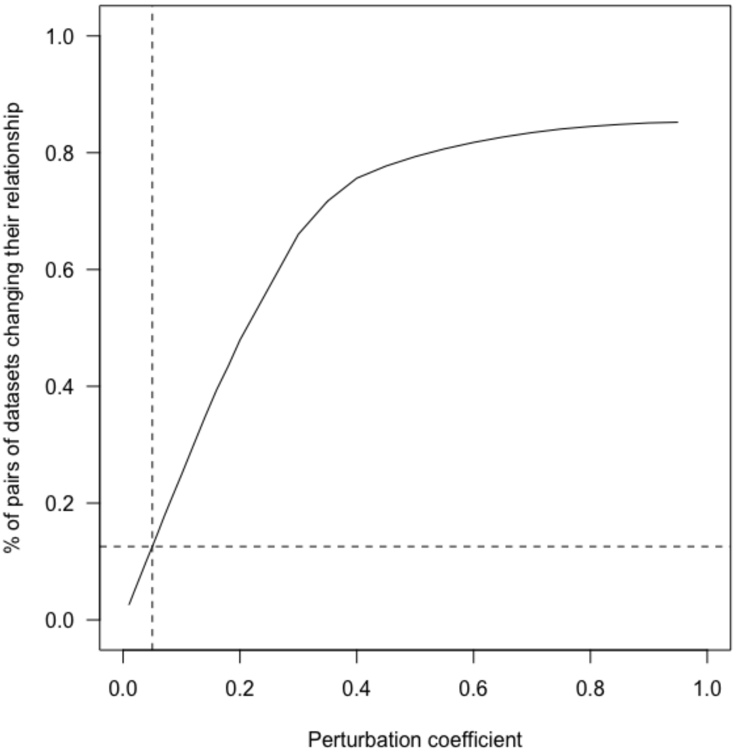


Figure 4.14: The percentage of dataset pairs reversing their order vs. the maximal percentage of the perturbation in the coefficients of the composite indicator.

CHAPTER 5

DISCUSSION

In the first part of this chapter, the results of the research are presented in the context of the research objectives, research question, and hypothesis that have been set. In the second part of this chapter, a brief review of existing studies that are most closely related to the efforts in this research is provided. In the third part of this chapter, some of the limitations of the conducted research are mentioned.

5.1 Results contribute to research objectives, research questions, and hypothesis

In accordance with the identified research problem, which concerns the questionable quality of open (meta)data, five research objectives (ROs) have been formulated at the beginning of this thesis. Furthermore, three research questions (RQs) have been defined to assist the author in carrying out this research. Additionally, a hypothesis (H) has been posited, relating to the operationalization of the theoretical framework for metadata quality for open datasets.

5.1.1 First research objective

The first and second research questions (RQ1, RQ2) are related to the first research objective (RO1):

RO1: To synthesize the results of previous research on the subject of quality of open (meta)data and dimensions identified for the purpose of measuring them.

RQ1: What are the key quality dimensions of open data metadata?

RQ2: How to measure identified metadata quality dimensions?

Brief answers to both questions are given below, based on the results obtained within the activity Explicate problem (subchapter 4.1.1, page 50). The answer to the first research question (RQ1), concerning the key quality dimensions of open data metadata, includes findability, retrievability, interoperability, reusability, and contextuality with respect to the properties of datasets, as well as completeness, conformance, coherence, accuracy, openness, retrievability, understandability, and timeliness with respect to the properties of metadata. Given that the research is based on the methodology for constructing composite indicators, in which the theoretical framework for a composite indicator is represented by a hierarchical structure of concepts, the initial matrix structure that emerges from the literature review is transformed into a hierarchical one. At the top of this structure, the composite indicator is positioned, beneath which elements are distributed across one or more levels, with individual indicators situated at the bottom (as described in subchapter 3.2.1, page 27). It is considered that each dimension, in terms of the metadata properties, should be assessed directly by one individual indicator, thus these dimensions are positioned immediately above the individual indicators within the structure. Furthermore, dimensions regarding the dataset properties, which are to be assessed by a group of individual indicators, are positioned above the dimensions related to the metadata properties. In addressing the second research question (RQ2) regarding the measurement of identified metadata quality dimensions, measurements are conducted using quality indicators for the metadata of open datasets. Initially, 97 individual indicators are defined, from which 71 are retained after excluding those that are semantically similar. In order for the previously defined objective (RO1) to be achieved:

- A literature review is conducted (see subchapter 1.1, chapter 2, subchapter 4.1.1);
- Metadata fields from different metadata standards are analysed and harmonised (subchapter 2.2.3).

By clarifying the observed problem and proposing the initial version of the theoretical framework (Šlibar, 2024c), RO1 was achieved within the first activity of the method framework for design science research.

5.1.2 Second research objective

Comprehension of the multidimensional phenomenon to be measured, the structure of the initial theoretical framework, elements of the initial theoretical framework, and the selection criteria for individual indicators served as the basis for defining the final theoretical framework of metadata quality for open datasets, which was established as the second research objective (RO2):

RO2: To define a theoretical framework of metadata quality for open datasets.

In order for the previously defined objective (RO2) to be achieved:

- Expert opinions are researched by combining the Q methodology with Lawshe's content validity ratio (subchapter 4.1.2).

By testing the content validity of the initial theoretical framework, 32 individual indicators for metadata quality of open datasets are found to be relevant (see Table 4.6). These relevant indicators are categorised into eight dimensions based on metadata properties and five dimensions based on dataset properties (see Figure 4.2). This fulfils the second research objective within the Define requirements activity.

5.1.3 Third research objective

The selected portals (i.e., open data portals from which open datasets are retrieved), the determined sample size (i.e., the number of open datasets for which the metadata quality are calculated), and the chosen relevant individual indicators form the basis for data collection, which was defined as the third research objective (RO3):

RO3: To collect and organize data about metadata from open data portals.

To ensure that previously defined objective (RO3) was achieved:

- Two open dataset portals are selected from which metadata about open datasets are retrieved, the official portal for European data and the central source of Australian open government data (see subchapter 3.1);
- The smallest sample size required to obtain a margin of error of at most ±2% at a 95% confidence level is determined, which is 2,401 datasets from each selected OGD portal (see subchapter 3.1);
- Metadata fields of each selected OGD portal are mapped with the metadata fields listed in (Šlibar, 2024b) and the obtained results are displayed in Table 4.7.
- The scores for the relevant individual indicators of each retrieved dataset are calculated and stored in the Harvard Dataverse research data repository (Šlibar, 2024a).

Metadata for 4,820 open datasets from two open government data portals (European and Australian) have been successfully retrieved, the metadata fields of each selected OGD portal have been mapped with the metadata fields listed in (Šlibar, 2024b), and the scores for the relevant individual indicators of 4,820 datasets have been calculated (Šlibar, 2024a). Thus, within the activity Design and develop artefact, the third research objective has been achieved (see subchapter 4.2.1, page 62).

5.1.4 Fourth research objective

The theoretical framework of metadata quality for open datasets was used as the basis for the development of a composite indicator of metadata quality for open datasets, which was established as the fourth research objective (RO4):

RO4: To define the composite indicator of metadata quality of open datasets.

In order for the previously defined objective (RO4) to be achieved:

- An appropriate method for multivariate data analysis has been selected, the principal components analysis (see 3.2.3);
- The weights of the dimensions and individual indicators have been calculated based on expert judgement (see Table 4.13, Table 4.24, Table 4.25, Table 4.26, Table 4.27, Table 4.28);

- A suitable method for the aggregation of elements of the theoretical framework has been chosen, which is the linear aggregation (see 3.2.3).

Thus, based on the previously described steps and their results, the fourth research objective has been achieved (see 4.2.1, page 62).

5.1.5 Fifth research objective

The composite indicator developed, based on the theoretical framework of metadata quality for open datasets, together with the collected and organised data, serves as the basis for the calculation of the composite indicator's values. This approach was established as the fifth research objective (RO5) and assists in providing answers to the third research question (RQ3):

RO5: To compute the values of the composite indicator on the collected data.

RQ3: How to assess the metadata quality?

Although initially more datasets were retrieved from the two selected portals (a total of 4,820), further examination led to the identification and removal of duplicates before conducting a multivariate data analysis. The composite scores, or rather, the scores of aggregates at all levels (dimensions and the composite indicator) were calculated for 4,622 open datasets. The computed scores for all observed datasets have been stored as a dataset on the Harvard Dataverse (Šlibar, 2024a). Thus, within the Demonstrate artefact activity, the fifth research objective has been achieved (see 4.2.2, page 77). Moreover, based on the results of this activity, an answer to the third research question (RQ3) is provided: The quality of metadata can be assessed through the application of a composite indicator to the metadata of collected open datasets.

In addition to the research objectives and questions, the author formulated a hypothesis regarding the developed composite indicator of the metadata quality of open datasets and undertook steps whose results supported this hypothesis (H1):

H1: The developed metadata quality composite indicator for open datasets is robust.

The evaluation of the robustness of the developed metadata quality composite indicator for open datasets pertains to the application of uncertainty and sensitivity analyses (see subchapter 3.2.5, page 45 and subchapter 4.2.3, page 82).

The sensitivity analysis, or, more precisely, Sobol' sensitivity measures showed that the developed composite indicator is more sensitive to the choice of aggregation method than to the perturbation of the weights of composite indicator elements across all levels in the structure (see Table 4.33). Although the choice of the aggregation method changes the composite indicator scores of the same dataset in the different replications (see Figure 4.9), it has no significant influence on the relationship between two different datasets, as the ranking of the datasets has remained the same in different replications (see Figure 4.10). This is also supported by testing the correlation coefficient based on Fisher's z-transformation at the statistical significance level

$p < 0.05$. The result of the one-tailed z-test showed that the Pearson correlation coefficient between the composite indicator scores or ranks, derived by geometric and linear aggregation, is statistically significantly greater than 0.99 at the level of statistical significance $p < 0.05$ (subchapter 4.2.3, page 82).

Since the choice of aggregation method in the developed composite indicator had no influence on the final ranking of the datasets, the robustness of the composite indicator was analysed with regard to the perturbation of the weights of composite indicator elements across all levels in the structure (the 4th uncertainty analysis). The results showed that even with an almost maximum weighting perturbation coefficient of 0.95, the ranking of 14.78% of dataset pairs remained consistent (see Figure 4.14).

5.2 Efforts to improve the quality of open data

To situate this research, or more precisely its results, within the broader academic conversation and to highlight its relevance and contribution to the field, a concise review of prior studies that are closely aligned with the current research's focus and efforts is provided below. Most closely related to the efforts in this research are Neumaier et al. (2016); Reiche and Höfig (2013); Ochoa and Duval (2009); Király (2019); Assaf et al. (2015); Milić et al. (2018); Consortium of data.europa.eu (n.d.).

The author has become aware that there is a need for harmonisation of different metadata schemas based on works by Assaf et al. (2015); Neumaier et al. (2016); Milić et al. (2018). Initially, the author has been focused on extending previous works done by Assaf et al. (2015); Neumaier et al. (2016); Milić et al. (2018). Therefore, the author tried to map the metadata fields of the different data management systems. However, during the mapping process, shortcomings of this approach were identified (e.g., the same metadata fields may have different properties/keys in different portals based on the same data management system), as detailed described at the end of subchapter 2.1 (page 18). Due to the noticed issues of this approach, a harmonisation of metadata fields was undertaken according to the metadata standards, which are recognised as prominent in the field of open data (Šlibar, 2024b). Another aspect lacking in works by Assaf et al. (2015); Neumaier et al. (2016); Milić et al. (2018) is an explicit definition of each metadata field, which can very easily lead to inconsistent metadata field usage. Thus, within this research, the definitions of metadata fields are provided, along with their mapping (Šlibar, 2024b).

Existing studies on the quality of metadata in open data do not make a clear distinction between quality dimensions in terms of metadata properties and dataset properties (Neumaier et al., 2016; Király, 2019; Reiche & Höfig, 2013; Kubler et al., 2018; Vetrò et al., 2016). A concern that arises is whether it is appropriate to assess quality merely for the sake of assessment or for another reason. For instance, the presence of an id, title, or tags influences the findability of an open dataset, whereas the presence of a license does not. This raises the question: 'Should the presence of the license be assessed, and if so, why?'. Consequently, this research has made a clear distinction between these two perspectives on metadata quality assessment and has contributed to a better understanding of the quality dimensions of open datasets metadata.

Reiche and Höfig (2013); Neumaier et al. (2016) have influenced the metadata quality indicators and dimensions, i.e., the development of the theoretical framework for metadata quality in this research. Although Reiche and Höfig (2013); Neumaier et al. (2016) deemed the 'timeliness' dimension relevant for the quality assessment of open dataset metadata, it was not incorporated into their framework. Hence, this research extends their work. Furthermore, although in existing studies on the automated quality assessment of metadata, the quality indicator 'richness of information / conformance to expectations' is considered important (Neumaier et al., 2016; Reiche & Höfig, 2013; Király, 2019; Ochoa & Duval, 2009), it was not included in this research. The author deemed it inappropriate because each time a new metadata record (in this research, an open dataset) is added to the repository, the metadata quality of all records should be recalculated. Therefore, a quality indicator that depends in any way on other indicators or new records should not be taken into account.

In parallel with the development of a composite indicator within this research, the Metadata Quality Assessment (MQA)[1] tool has been developed by the consortium of data.europa.eu. The MQA tool has greatly influenced the definition of dimensions within this research. Similarly to the composite indicator developed in this research, the MQA tool is based on quality indicators designed for automated assessment (Consortium of data.europa.eu, n.d.). Currently, MQA provides a breakdown of the scores that each portal has achieved across various dimensions, as well as the total scores awarded to each portal. At the dataset level, it offers insights into the presence of metadata for each quality indicator. However, unlike in this research (see Figure 4.8 or Table 4.30), MQA's indicators have not been used to score or rate the datasets individually.

5.3 Limitations of the research

Despite the fact that the research is carried out successfully and complies with the criteria set for ensuring the scientific contribution of the construction of the composite indicators and the greater practical relevance of the scientific results produced under the DS and DSR in IS discipline, there are some limitations that should be considered. The next paragraphs address the limitations with the respect to the five main activities of the method framework for design science research.

As part of the first activity Explicate problem, key criteria for the design and development of the artefact, i.e. a composite indicator based on the metadata quality framework for open datasets, were defined. Thus, one criterion specifies that the evaluation of metadata quality must rely solely on metadata automatically read from the portals and publicly available information. However, this requirement also imposes a limitation on the research. This means that it is not possible to verify the accuracy of all metadata contents. For instance, although there may be data, such as an email address, provided for contacting someone for more information about

[1] `https://data.europa.eu/mqa/?locale=en`.

the dataset, it could be recorded in the correct format without guaranteeing whether that person still holds that responsibility, perhaps having moved from the department or organisation. Both the processes (the methods or approaches used) and the outcomes (the results or conclusions derived) within the activity Explicate problem are, to some extent, subjective and influenced by the knowledge and skills of the author. For example, some relevant sources that could be used to establish a general metadata schema for open datasets and their distributions might be unintentionally overlooked in the literature search. The process of harmonising metadata properties/attributes of the metadata standard (that have been identified as relevant) is subject to possible misinterpretation, as harmonisation is done by mapping attributes based on the correspondence of their semantic meaning.

Although experts from different stakeholder groups are involved in the activity Define requirements, some experts did not choose one main role that they have in the context of open data (as they were asked to), but several. This certainly does not affect one of the main objectives of the Q methodology, which is to identify a pattern of opinion, i.e., whether there is one or more patterns. However, the choice of more than one role can easily affect data interpretation, especially when more than one pattern is identified, and one wants to determine whether experts belonging to the same group of stakeholders hold a similar opinion. The expertise of those who have chosen to participate is also reflected to some extent in the research findings. Due to the anonymisation of personal data, there is a low probability that one of the invited experts submitted his or her results more than once.

As with the first activity Explicate problem, the activity Design and develop artefact could be influenced by the author's subjectivity, knowledge, experience and skills. For instance, the mapping of individual portal properties to the established general metadata schema for open datasets and their distributions is done by the author based on the correspondence of their semantic meaning (Šlibar, 2024b). Another limitation of this activity is that there could be some errors in the R scripts for retrieving data from selected portals, pre-processing these data and analysing the processed data using appropriate research methods. It is also possible that the APIs of the portals to which the scripts connect provide incorrect data or have limited availability during the retrieval of the data. Another limitation is that in both selected portals, the European OGD portal and the Australian OGD portal, some individual indicators did not show sufficient variability, which affected the multivariate statistics, and such individual indicators could not be included in the multivariate data analysis. Therefore, it remains uncertain how these individual indicators will perform on other portals.

Like other activities in the design science research method framework, the activity Demonstrate artefact is influenced by the results of the previous activities. Considering that this research is limited to open government data only, the developed composite indicator is applied to datasets retrieved from portals where data is published by the public sector.

The evaluation of the artefact involved assessing the robustness of the developed composite indicator through uncertainty and sensitivity analyses. This approach might only be partially sufficient to determine the extent to which the artefact solves or mitigates the real problem that initiated this research. The evaluation did not encompass an assessment from the perspective of criterion validity. Thus, the associations between the composite indicator of metadata quality of open datasets and characteristics of public administration (e.g., transparency, openness of the public sector, citizen engagement), for which corresponding and valid indicators exist, were not examined. This indicates a gap in the evaluation process of the artefact, highlighting aspects that were not covered but are considered important for a comprehensive understanding of the artefact's effectiveness and relevance to public administration.

R scripts are used to analyse data with suitable research methods in all stages of the design science research method framework, except for the initial stage. Therefore, there could be some errors.

CHAPTER 6

CONCLUSION

This chapter summarizes research results by placing them in the context of their contributions. Given that this research is based on a combination of two methodologies, the design science research method framework and the methodology for constructing composite indicators, in order to ensure the scientific contribution of the construction of composite indicators as well as to meet the requirements for greater practical relevance of scientific results under the DS and DSR in information systems, the research contributes to both the scientific and practical communities. Therefore, the scientific and practical contributions are highlighted in the following subchapters. Furthermore, recommendations for further research are briefly presented in this concluding chapter.

6.1 Scientific contributions

In accordance with the defined research objectives and hypothesis, the following scientific contributions are reached:

1. Systematization and synthesis of existing knowledge in the domain of the quality of open (meta)data and dimensions identified for the purpose of measuring them.

   This scientific contribution was achieved through a literature review. The research is based on two methodological theories: the design science research method framework and the methodology for constructing composite indicators (as described in subchapter 1.3, page 8). Furthermore, existing literature was used to define open data and their infrastructure, focusing on open data management systems (see subchapter 1.1, page 4, the introduction to chapter 2, page 15, and subchapter 2.1, page 18), to analyse and harmonise metadata fields across different metadata standards (see subchapter 2.2.3, page 23), to define the concept of metadata quality (see subchapter 1.1, page 4 and subchapter 2.2.4, page 24), and to identify quality indicators for the metadata of open datasets (see subchapter 4.1, page 50).

   In this way, the second requirement posed by Johannesson and Perjons (2014) was met: to ensure the production of well-established and original results, these results need to be related to an existing knowledge base.

2. The development of a theoretical framework of metadata quality for open datasets.

   This scientific contribution has been realised by developing a theoretical framework, following activities within the design science research method framework and steps within the methodology for constructing composite indicators (see subchapter 4.1, page 50).

The development of a theoretical framework, which is a hierarchical structure of concepts for the composite indicator, involved a systematic literature review (see subchapter 3.2.1, page 27 and subchapter 4.1.1, page 50), the analysis and mapping of metadata fields from various international standards and specifications for open data management systems (see chapter 2, page 15), and expert opinion research (see subchapter 3.2.2, page 28 and subchapter 4.1.2, page 56).

3. The development of a composite indicator of metadata quality for open datasets.

   This scientific contribution was achieved by constructing and validating a composite indicator, following activities within the design science research method framework and steps in the methodology for constructing composite indicators (see subchapter 4.2, page 61).

   The construction of the composite indicator of metadata quality for open datasets was based on a theoretical framework and included: retrieving metadata of open datasets from open data portals that are based on different data management systems, mapping the metadata fields retrieved from selected portals to the fields defined in relevant metadata standards, calculating the scores of relevant individual indicators for all retrieved open datasets, analysing these scores with multivariate analysis, determining the importance/weights of individual indicators and dimensions using the analytic hierarchy process, and aggregating the scores of individual indicators and dimensions into a composite indicator score using linear aggregation (see subchapter 3.2.3, page 31 and subchapter 4.2.1, page 62).

   The composite indicator was validated through the evaluation of the robustness of the developed composite indicator. This involved the application of uncertainty and sensitivity analyses (see subchapter 3.2.5, page 45 and subchapter 4.2.3, page 82).

4. Results of the empirical research on metadata quality for open data.

   This scientific contribution has been realised through the application of a developed composite indicator to a large random sample of open datasets retrieved from two open data portals, which are based on different data management systems. Some of the results of the empirical research are listed below. Certain metadata fields of the relevant individual indicators are not found on either of the portals (see Table 4.7). Descriptive statistics of dichotomous and continuous individual indicators have revealed that some individual indicators exhibit no variability across datasets, or that some individual indicators yield identical scores in all the observed datasets (see Table 4.8, Table 4.9). It has also been shown by empirical research that Australian datasets achieve higher scores or have better metadata quality across all aggregates compared to European datasets, with the exception of the reusability dimension, where European datasets are slightly better than Australian datasets (as visualised in Figure 4.7). Although the interoperability dimension is assigned a lower weight / aggregated priority compared to other dimensions (see Table 4.13), it has been found to contribute to the variability of the composite scores on both the Australian OGD portal and the European OGD portal, given the high standard deviation (as can be seen in Figure 4.7).

As already explained, this research is based on the combination of two relevant methodologies: the method framework for design science research (Johannesson & Perjons, 2014) and the methodology for constructing composite indicators (OECD et al., 2008). The development of a theoretical framework and a composite indicator of metadata quality for open datasets is undertaken through five main activities of the method framework for design science research (considered as the outer cycle within this research) and ten steps of the methodology for constructing composite indicators (considered as the inner cycle within this research). These methodologies have guided the author in selecting appropriate research methods for each activity and the steps within each activity, as can be seen in subchapter 3.2 (page 27). In this way, the first requirement proposed by Johannesson and Perjons (2014) was met, stipulating that for the creation of new knowledge relevant to global practice, rigorous research methods must be applied in the research.

To gather expert opinions on the relevance of individual indicators for specific quality dimensions of open data metadata (as explained in subchapter 3.2.2, page 28), as well as their opinions on the relative importance of elements within the developed theoretical framework (as explained in subchapter 3.2.3, page 31), findings from earlier activities were shared with experts. This approach was critical to ensuring that the experts comprehensively understood the tasks at hand, enabling them to provide informed feedback. As a result, their contributions were instrumental in developing both the theoretical framework and the composite indicator of metadata quality for open datasets. The dissertation will be accessible after the defence through the digital repository of the Faculty of Organization and Informatics[1]. Additionally, it is planned to disseminate findings of this research through publications at international scientific conferences and in scientific journals. In this way, the third requirement proposed by Johannesson and Perjons (2014), to share the results obtained with practitioners and researchers, was partially fulfilled.

6.2 Practical contributions

It is important to mention that two different perspectives on metadata quality assessment were considered when building the theoretical framework (as shown in subchapter 4.1.1, page 50). One perspective, more advocated by the practical community, is where the quality of metadata for open datasets is assessed in terms of the dataset properties, while the other, more supported by the academic community, is where the quality of metadata for open datasets is assessed in terms of the properties of the metadata itself.

For OD/IT strategists within the public sector, OD advisors in the public sector and external

[1] Link to the digital repository of the Faculty of Organization and Informatics: `https://repozitorij.foi.unizg.hr/en`

OD consultants, this is a call to develop strategies and action plans that promote best practise in publishing datasets as open data. It is also important that those responsible for the implementation and maintenance of data portals provide clear guidance to OD generators in the public sector on the publication of data, introduce a quality control system or, even better, introduce and enforce quality assurance of metadata. Also, it should be considered by OD/IT strategists within the public sector, OD advisors in the public sector and external OD consultants that for certain individual indicators, deemed relevant by experts, the necessary metadata fields were absent in a very large random sample of open datasets retrieved from the European and Australian OGD portals.

This research aims to raise awareness and sensitise all stakeholders involved with open data, including OD generators in the public sector, OD users, OD activists, OD/IT strategists within the public sector, OD advisors in the public sector, external OD consultants, and politicians, to the fact that the quality of metadata for open datasets is diverse and that few datasets have achieved a higher composite indicator score. According to the developed composite indicator, each dataset can achieve scores from 0 to 1. It has been observed that the composite scores of the datasets retrieved from the European OGD portal range between 0.182 and 0.585 (see Table 4.32), whilst those from the Australian OGD portal range between 0.343 and 0.869 (see Table 4.31). This indicates that none of the selected datasets, retrieved from these two portals, has achieved the maximum score of the composite indicator. Furthermore, the open datasets achieve varying scores at the dimension level. The Australian datasets, however, all received identical scores in the reusability dimension because the individual indicators, which are part of the reusability dimension, lacked metadata for the datasets retrieved from the Australian OGD portal (see Table 4.31). It is also worth mentioning that 25% of the datasets from the European OGD portal achieved a composite indicator score higher than 0.436, whereas 25% of those retrieved from the Australian OGD portal achieved a composite indicator score higher than 0.735.

The composite indicator that has been developed is found to be useful for benchmarking. It has been developed for assessing the metadata quality of open datasets and for comparing open datasets against each other based on the calculated values of individual indicators and aggregate values (dimensions and composite indicator). Thus, the developed composite indicator has been applied to a random sample of datasets retrieved from two open data portals (the Australian OGD portal and the European OGD portal), and the results have been stored as a dataset in the Harvard Dataverse research data repository (Šlibar, 2024a). Moreover, the developed composite indicator can easily be adapted for comparing open data portals or even countries. One approach for achieving this would be to take random samples of open datasets from portals or countries. Then, these datasets should be assessed by calculating scores for individual indicators, dimensions, and the composite indicator. Following this, the scores of the composite indicator should be aggregated by computing statistical measures such as the arithmetic mean, median, or another relevant summary measure, focusing on either the portal or country level for analysis. This exact approach has been employed (as can be seen in subchapter 4.2.2, page 77) for the comparison of the Australian OGD portal and the European OGD portal.

The theoretical framework and the developed composite indicator can be used and applied by various interest groups for diverse purposes. On open data portals, the scores for individual indicators, dimensions, and the composite indicator of datasets could be published alongside the datasets themselves, or alternatively, only the scores for dimensions might be presented using radial plots (as demonstrated in subchapter 4.2.2, Figure 4.8). This information can be of value to both data providers and data consumers.

6.3 Recommendations for future research

In further research, it is planned to convert the developed composite indicator into an interactive web application using the R package Shiny[2], which will potentially accelerate the process of disseminating the results, i.e., a new artefact, as well as the knowledge about this artefact and its impact on the environment.

Although only open government data is the focus of this research, the individual indicators in the final theoretical framework presented in subchapter 4.1.2 (page 56) are clearly general enough to be applicable to other open data. Therefore, one of the recommendations for the future relates to the application of the results to open data that does not come from the public sector, but from other institutions, companies, or individuals. Furthermore, it would be of great interest to apply the methodology of this research not only to open datasets, but also to non-open datasets, e.g. within a closed system such as a large private company that shares data internally.

Another important recommendation relates to the theoretical framework developed and its operationalisation. Since the information infrastructure is a prerequisite for the existence of open data (as already elaborated in subchapter 1.1, page 4 and chapter 2, page 15), the theoretical framework developed, and thus the composite indicator, needs to be updated in light of the changing nature of the information technology. The need for a change may also arise due to changes in legislation, e.g., a new directive having to replace the one currently in force.

Additionally, it has been indicated by the research that different patterns of thought regarding the relevance of individual indicators within certain dimensions (contextuality, interoperability, and reusability) have been observed among experts. However, the level of agreement within each profile, as determined by the number of experts concurring, has been found to be insufficient. Thus, the identification of relevant individual indicators for each profile has been prevented by the lack of consensus. Consequently, future research could be undertaken with a larger group of experts. This would aid in ascertaining whether the development of separate composite indicators for each expert profile would be beneficial.

Since the focus of this research is not on the architecture of information systems but on the metadata of open datasets, the chapter 2 (page 15) provides an initial mapping of concepts

[2] `https://shiny.posit.co/`

commonly used in this emerging area of open data research to those specified in OAIS Reference Model (Consultative Committee for Space Data Systems, 2019). Therefore, this first version can be extended in future works, especially in the studies that will deal with architecture itself.

Further research could include the evaluation of the artefact, i.e., the composite indicator based on the metadata quality framework for open datasets, from the aspect of criterion validity. As discussed in chapter 2 (page 15), OGD can have an impact on other public administration characteristics, such as the potential for innovation, increasing the transparency and openness of the public sector, engaging citizens, and enabling better law enforcement. As the computed scores of the composite indicator assigned to the datasets can be easily aggregated, different portals or countries can be compared. It can thus be assumed that an individual country has higher quality metadata if it has a more transparent and open public administration, more advanced e-government, greater citizen participation and greater capacity for innovation. There are already indicators for these characteristics of public administration. For example, the indicators eGovernment Benchmark - European Commission, Sustainable Governance Indicators - Bertelsmann Stiftung, Corruption Perception Index - Transparency International assess the transparency and openness of public administration (Palaric, Thijs, Hammerschmid, & Directorate-General for Employment, Social Affairs and Inclusion of the EC, 2018; Van Dooren & Directorate-General for Employment, Social Affairs and Inclusion of the EC, 2018). Therefore, the associations between the composite indicator of metadata quality of open datasets and the external behaviour, i.e., public administration characteristics for which there will be corresponding and valid indicators at that time, can be examined.

BIBLIOGRAPHY

Aczél, J., & Saaty, T. L. (1983, March). Procedures for synthesizing ratio judgements. *Journal of Mathematical Psychology*, *27*(1), 93–102. Retrieved 2024-01-05, from `https://www.sciencedirect.com/science/article/pii/0022249683900287` doi: 10.1016/0022-2496(83)90028-7

Adler, D., Kelly, S. T., Elliott, T. M., & Adamson, J. (2022, December). *vioplot: Violin Plot.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/vioplot/index.html`

Ali, M., Alexopoulos, C., & Charalabidis, Y. (2022, November). A comprehensive review of open data platforms, prevalent technologies, and functionalities. In *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance* (pp. 203–214). Guimarães, Portugal: Association for Computing Machinery. doi: 10.1145/3560107.3560142

Allen, M. (Ed.). (2017a). *Validity, Concurrent* (4th ed.). Thousand Oaks, California: SAGE Publications, Inc. Retrieved from `https://methods.sagepub.com/reference/the-sage-encyclopedia-of-communication-research-methods` doi: 10.4135/9781483381411

Allen, M. (Ed.). (2017b). *Validity, Face and Content* (4th ed.). Thousand Oaks, California: SAGE Publications, Inc. Retrieved from `https://methods.sagepub.com/reference/the-sage-encyclopedia-of-communication-research-methods` doi: 10.4135/9781483381411

Aoki, H., Kim, J., & Lee, W. (2013, January). Propagation & level: Factors influencing in the ICT composite index at the school level. *Computers & Education*, *60*(1), 310–324. doi: 10.1016/j.compedu.2012.07.013

Assaf, A., Troncy, R., & Senart, A. (2015). HDL - Towards a harmonized dataset model for open data portals. In *Joint Proceedings of USEWOD and PROFILES 2015* (Vol. 1362). Portoroz, SLOVENIA. Retrieved from `https://ceur-ws.org/Vol-1362/`

Association for Computing Machinery. (2020, August). *Artifact Review and Badging - Current: Artifact Review and Badging Version 1.1.* Retrieved 2022-11-14, from `https://www.acm.org/publications/policies/artifact-review-and-badging-current`

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015, October). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. Retrieved 2024-01-01, from `https://www.sciencedirect.com/science/article/pii/S0740624X1500091X` doi: 10.1016/j.giq.2015.07.006

Ayre, C., & Scally, A. J. (2014, January). Critical Values for Lawshe's Content Validity Ratio: Revisiting the Original Methods of Calculation. *Measurement and Evaluation in*

Counseling and Development, *47*(1), 79–86. doi: 10.1177/0748175613513808

Banasick, S. (2023a, September). *EQ Web Configurator.* Retrieved 2024-01-05, from `https://github.com/shawnbanasick/eq_web_configurator` doi: 10.5281/zenodo.8337126

Banasick, S. (2023b, September). *EQ Web Sort.* Retrieved 2024-01-05, from `https://github.com/shawnbanasick/eq-web-sort` doi: 10.5281/zenodo.8339819

Basak, I., & Saaty, T. L. (1993, February). Group decision making using the analytic hierarchy process. *Mathematical and Computer Modelling*, *17*(4), 101–109. Retrieved 2024-01-05, from `https://www.sciencedirect.com/science/article/pii/0895717793901793` doi: 10.1016/0895-7177(93)90179-3

Baskerville, R., Lyytinen, K., Sambamurthy, V., & Straub, D. (2011, January). A response to the design-oriented information systems research memorandum. *European Journal of Information Systems*, *20*(1), 11–15. doi: 10.1057/ejis.2010.56

Becker, W. (2023, October). *COINr: Composite Indicator Construction and Analysis.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/COINr/index.html`

Berends, J., Carrara, W., Engbers, W., Vollers, H., & Publications Office of the European Union. (2020). *Reusing open data: a study on companies transforming open data into economic and societal value* (Tech. Rep. No. OA-03-20-042-EN-N). Luxembourg: Publications Office of the European Union. doi: 10.2830/876679

Berends, J., Carrara, W., Vollers, H., & Publications Office of the European Union. (2020). *Barriers in working with open data* (Tech. Rep. No. OA-BF-20-005-EN-N). Luxembourg: Publications Office of the European Union. doi: 10.2830/88151

Bogdanović, M., Veljković, N., Frtunić Gligorijević, M., Puflović, D., & Stoimenov, L. (2021, January). On revealing shared conceptualization among open datasets. *Journal of Web Semantics*, *66*. Retrieved 2023-11-08, from `https://www.sciencedirect.com/science/article/pii/S1570826820300573` doi: 10.1016/j.websem.2020.100624

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012, January). The State of Open Data: Limits of Current Open Data Platforms. In *Proceedings of the 21st International Conference on World Wide Web.* Lyon, France: Association for Computing Machinery, Inc.

Brown, S. R. (1993, April). A Primer on Q Methodology. *Operant Subjectivity*, *16*(3/4), 91–138.

Bruce, T. R., & Hillmann, D. I. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In *Metadata in Practice.* Chicago: ALA Editions. Retrieved 2024-01-03, from `https://hdl.handle.net/1813/7895`

Chatfield, C., & Collins, A. J. (1980). *Introduction to Multivariate Analysis* (First edition ed.). CHAPMAN & HALL/CRC. doi: 10.1007/978-1-4899-3184-9

Conboy, K., Fitzgerald, G., & Mathiassen, L. (2012, March). Qualitative methods research in information systems: motivations, themes, and contributions. *European Journal of*

Information Systems, *21*(2), 113–118. doi: 10.1057/ejis.2011.57

Consortium of data.europa.eu. (n.d.). *Metadata Quality Assessment Methodology* (Tech. Rep.). Retrieved 2023-12-26, from `https://data.europa.eu/mqa/methodology?locale=en`

Consultative Committee for Space Data Systems. (2019, December). *REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS) PROPOSED DRAFT RECOMMENDED PRACTICE.* CCSDS Secretariat, National Aeronautics and Space Administration.

Couture-Beil, A. (2022, January). *rjson: JSON for R.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/rjson/index.html`

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (4th edition ed.). Los Angeles, California: SAGE Publications, Inc.

Daga, E., Panziera, L., & Pedrinaci, C. (2015). A BASILar approach for building web APIs on top of SPARQL endpoints. In *Proceedings of the Third Workshop on Services and Applications over Linked APIs and Data* (Vol. 1359, pp. 22–32).

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019, April). *xtable: Export Tables to LaTeX or HTML.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/xtable/index.html`

DAMA International. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2nd Edition ed.). Basking Ridge, New Jersey: Technics Publications.

Data.gov admins. (2013, April). *Open Data: A History.* Retrieved 2024-01-01, from `https://data.gov/blog/open-data-history/`

Davies, T., Walker, S. B., Rubinstein, M., & Perini, F. (Eds.). (2019). *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and the nternational Development Research Centre.

Dawes, S. S., & Helbig, N. (2010). Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency. In *Electronic Government* (pp. 50–60). Lausanne, Switzerland: Springer. doi: 10.1007/978-3-642-14799-9_5

DCMI Usage Board. (2012, June). *Dublin Core™ Metadata Element Set, Version 1.1: Reference Description.* Dublin Core Metadata Initiative. Retrieved 2024-01-05, from `https://www.dublincore.org/specifications/dublin-core/dces/`

Dekkers, M., Loutas, N., De Keyzer, M., & Goedertier, S. (2012). *Open Data & Metadata Quality | Training Module 2.2.* Retrieved 2024-01-03, from `https://joinup.ec.europa.eu/collection/open-government/document/tm22-open-data-metadata-quality-en`

Deutz, D. B., Buss, M. C. H., Hansen, J. S., Hansen, K. K., Kjelmann, K. G., Larsen, A. V., ... Holmstrand, K. F. (2020, June). *How to FAIR: a website to guide researchers on making research data more FAIR.* Zenodo. doi: 10.5281/zenodo.3712065

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., & Zijlstra, T. (n.d.). *API.* Retrieved 2024-01-04, from https://opendatahandbook.org/glossary/en/terms/api/

Directorate-General for Communications Networks, Content and Technology of the EC, & EC. (n.d.-a). *European legislation on open data.* Retrieved 2024-01-01, from https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

Directorate-General for Communications Networks, Content and Technology of the EC, & EC. (n.d.-b). *From the Public Sector Information (PSI) Directive to the Open Data Directive.* Retrieved 2023-09-03, from https://digital-strategy.ec.europa.eu/en/policies/psi-open-data

Dresch, A., Lacerda, D. P., & Antunes Júnior, A. V. (2014). *Design Science Research: A Method for Science and Technology Advancement* (1st ed.). Springer Cham.

EC, & Directorate-General for Employment, Social Affairs and Inclusion of the EC. (2010, December). *COMMISSION STAFF WORKING DOCUMENT: Free movement of workers in the public sector.* Retrieved 2024-01-19, from https://ec.europa.eu/social/main.jsp?langId=en&catId=465

Epskamp, S., Costantini, G., Haslbeck, J., & Isvoranu, A. (2023, November). *qgraph: Graph Plotting Methods, Psychometric Data Visualization and Graphical Model Estimation.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/qgraph/index.html

Ermis-Demirtas, H. (2018, December). Establishing Content-Related Validity Evidence for Assessments in Counseling: Application of a Sequential Mixed-Method Approach. *International Journal for the Advancement of Counselling*, *40*(4), 387–397. doi: 10.1007/s10447-018-9332-4

Farrow, R., Iniesto, F., Weller, M., & Pitt, R. (2020). *GO-GN Research Methods Handbook.* Milton Keynes: Global OER Gradate Network. Retrieved 2024-01-03, from http://go-gn.net/wp-content/uploads/2020/07/GO-GN-Research-Methods.pdf

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R* (1st edition ed.). SAGE Publications Ltd.

Fitzgerald, A. (2022, April). *What Is a Data Repository? [+ Examples and Tools].* Retrieved 2024-01-19, from https://blog.hubspot.com/website/data-repository

Fletcher, T. D. (2023, November). *psychometric: Applied Psychometric Theory.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/psychometric/index.html

Foster, J., McGillivray, M., & Seth, S. (2009, January). *Rank robustness of composite indices, OPHI Working Paper 26.* University of Oxford. Retrieved from https://ophi.org.uk/working-paper-number-26/

Francis, F. C., Haider, S., Foskett, D. J., & Estabrook, L. S. (n.d.). *Catalog standardization.* Retrieved 2024-01-05, from https://www.britannica.com/topic/library/

Catalog-standardization

Furrie, B. (2003). *Understanding Marc Bibliographic: Machine-Readable Cataloging* (7th edition ed.). Washington, DC: Library of Congress.

Gao, Y., Janssen, M., & Zhang, C. (2021). Understanding the evolution of open government data research: towards open data sustainability and smartness. *International Review of Administrative Sciences*, *89*(1), 59–75. doi: 10.1177/00208523211009955

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

González-Mora, C., Barros, C., Garrigós, I., Zubcoff, J., Lloret, E., & Mazón, J.-N. (2023, January). Improving open data web API documentation through interactivity and natural language generation. *Computer Standards & Interfaces*, *83*. Retrieved 2023-11-08, from https://www.sciencedirect.com/science/article/pii/S0920548922000344 doi: 10.1016/j.csi.2022.103657

Herzberg, B. (2022, January). *Metadata Repositories: Data Dictionary vs. Data Inventory vs. Data Catalog.* Retrieved 2024-01-19, from https://blog.satoricyber.com/metadata-repositories-data-dictionary-vs-data-inventory-vs-data-catalog/

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. Retrieved 2024-01-03, from https://www.jstor.org/stable/25148625 doi: 10.2307/25148625

Holmes, A., Illowsky, B., Dean, S., & Hadley, K. (2017). *Introductory Business Statistics* (First Edition ed.). Houston, Texas: XanEdu Publishing Inc.

Homma, T., & Saltelli, A. (1996, April). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, *52*(1), 1–17. Retrieved 2024-01-05, from https://www.sciencedirect.com/science/article/pii/0951832096000026 doi: 10.1016/0951-8320(96)00002-6

Horn, J. L. (1965, June). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi: 10.1007/BF02289447

Hornik, K., & Leisch, F. (2023, November). *tools: Tools for Package Development.* Retrieved 2024-01-05, from https://r-universe.dev/manuals/tools.html

Hunnius, S., Njacheun-Njanzoua, G., Prasopoulou, E., Hardinges, J., Fawcett, J., & Heath, T. (2015). *Open Data Stakeholder Requirement Report 2* (Project's work package No. D2.6).

Huyer, E., van Knippenberg, L., & Publications Office of the European Union. (2020). *The economic impact of open data – Opportunities for value creation in Europe* (Tech. Rep.). Publications Office. doi: 10.2830/63132

ISO. (2012). *ISO 14721:2012.* Geneva: ISO. Retrieved 2024-01-04, from https://www.iso.org/standard/57284.html

ISO. (2014). *ISO 19115-1:2014.* ISO. Retrieved 2024-06-14, from https://www.iso.org/

standard/53798.html

ISO. (2015). *ISO 9000:2015.* ISO. Retrieved 2024-06-14, from https://www.iso.org/obp/ui/en/#iso:std:iso:9000:ed-4:v1:en:term:3.6.2

ISO. (2019). *ISO 8601.* ISO. Retrieved 2024-01-05, from https://www.iso.org/iso-8601-date-and-time-format.html

ISO and IEC. (2013). *ISO/IEC 11179-3:2013.* Geneva: ISO. Retrieved 2023-12-26, from https://www.iso.org/standard/50340.html

ISO and IEC. (2015). *ISO/IEC 11179-6:2015.* Geneva: ISO. Retrieved 2023-12-26, from https://www.iso.org/standard/60342.html

ISO and IEC. (2019). *ISO/IEC 11179-7:2019.* Geneva: ISO. Retrieved 2023-12-26, from https://www.iso.org/standard/68766.html

ISO and IEC. (2023). *ISO/IEC 11179-33:2023.* Geneva: ISO. Retrieved 2023-12-26, from https://www.iso.org/standard/81725.html

Jacobson, D., Brail, G., & Woods, D. (2012). *APIs: A Strategy Guide: Creating Channels with Application Programming Interfaces* (1st edition ed.). O'Reilly Media.

James, D., & Hornik, K. (2023, May). *chron: Chronological Objects which Can Handle Dates and Times.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/chron/index.html

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012, September). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268. doi: 10.1080/10580530.2012.716740

Jethani, S., & Leorke, D. (2021). *Openness in Practice: Understanding Attitudes to Open Government Data* (First edition ed.). Palgrave Macmillan.

Johannesson, P., & Perjons, E. (2014). *An Introduction to Design Science.* Springer Publishing Company, Incorporated.

Juran, J. M., & DeFeo, J. A. (2010). *Juran's Quality Handbook: The Complete Guide to Performance Excellence* (6th edition ed.). New York: McGraw-Hill Companies.

Khder, M. A. (2021, December). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*, *13*(3), 145–168. Retrieved 2023-11-09, from http://ijasca.zuj.edu.jo/PapersUploaded/2021.3.11.pdf doi: 10.15849/IJASCA.211128.11

Király, P. (2019). *Measuring metadata quality* (PhD dissertation, University of Göttingen). Retrieved from https://ediss.uni-goettingen.de/handle/21.11130/00-1735-0000-0003-C17C-8

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, P. O., Turner, M., Niazi, M., & Linkman, S. (2010, August). Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology*, *52*(8), 792–805. Retrieved 2024-01-09, from https://www.sciencedirect.com/science/article/pii/S0950584910000467 doi: 10.1016/j.infsof.2010.03.006

Kubler, S., Jérérmy, R., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018, January). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, *35*(1), 13–29. Retrieved 2020-02-12, from https://linkinghub.elsevier.com/retrieve/pii/S0740624X16301319 doi: 10.1016/j.giq.2017.11.003

Kučera, J., & Chlapek, D. (2014). Benefits and Risks of Open Government Data. *Journal of Systems Integration*, *5*(1), 30–41. Retrieved 2023-09-07, from https://www.proquest.com/docview/1690665526/abstract/BD7F8352D0644D06PQ/1

Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open Government Data Catalogs: Current Approaches and Quality Perspective. In A. Kő, C. Leitner, H. Leitold, & A. Prosser (Eds.), *Technology-Enabled Innovation for Democracy, Government and Governance* (pp. 152–166). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-40160-2_13

Lavoie, B. (2014, October). *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)* (Tech. Rep.). Great Britain: Digital Preservation Coalition. Retrieved 2023-10-18, from http://www.dpconline.org/component/docman/doc_download/1359-dpctw14-02

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. doi: 10.1111/j.1744-6570.1975.tb01393.x

Lisowska, B. (2016, December). *Metadata for the open data portals* (Discussion Paper No. 6).

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic Reviews and Meta-Analysis* (1st edition ed.). Oxford: Oxford University Press.

Lourenço, R. P. (2015, July). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, *32*(3), 323–332. Retrieved 2024-01-01, from https://www.sciencedirect.com/science/article/pii/S0740624X15000660 doi: 10.1016/j.giq.2015.05.006

March, S. T., & Smith, G. F. (1995, December). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. doi: 10.1016/0167-9236(94)00041-2

Martin, A. D., Quinn, K. M., & Park, J. H. (2022, April). *MCMCpack: Markov Chain Monte Carlo (MCMC) Package.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/MCMCpack/index.html

Microsoft Corporation, & Weston, S. (2022a, February). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/doParallel/index.html

Microsoft Corporation, & Weston, S. (2022b, February). *foreach: Provides Foreach Looping Construct.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/foreach/index.html

Milić, P., Veljković, N., & Stoimenov, L. (2018). Comparative Analysis of Metadata Models on e-Government Open Data Platforms. *IEEE Transactions on Emerging Topics in*

Computing, *9*(1), 119–130. Retrieved 2023-11-08, from https://ieeexplore.ieee.org/document/8315058 doi: 10.1109/TETC.2018.2815591

Musa, A., Bebić, D., & Đurman, P. (2015, June). Transparency and Openness in Local Governance: A Case of Croatian Cities. *Hrvatska i komparativna javna uprava: časopis za teoriju i praksu javne uprave*, *15*(2), 415–450. Retrieved 2024-01-04, from https://hrcak.srce.hr/142730

Máchová, R., Hub, M., & Lněnička, M. (2018, January). Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective. *Aslib Journal of Information Management*, *70*(3), 252–268. doi: 10.1108/AJIM-02-2018-0026

Máchová, R., & Lněnička, M. (2017, January). Evaluating the Quality of Open Data Portals on the National Level. *Journal of Theoretical and Applied Electronic Commerce Research*, *12*(1), 21–41. doi: 10.4067/S0718-18762017000100003

Máchová, R., & Lněnička, M. (2019, January). A multi-criteria decision making model for the selection of open data management systems. *Electronic Government*, *15*(4), 372–391. doi: 10.1504/EG.2019.102579

Müller, K., & Wickham, H. (2023, March). *tibble: Simple Data Frames.* Retrieved 2024-01-09, from https://cran.r-project.org/web/packages/tibble/index.html

Nakazawa, M. (2024, January). *fmsb: Functions for Medical Statistics Book with some Demographic Data.* Retrieved 2024-02-27, from https://cran.r-project.org/web/packages/fmsb/index.html

Neumaier, S., Umbrich, J., & Polleres, A. (2016, October). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality*, *8*(1), 1–29. doi: 10.1145/2964909

Nikiforova, A., Bičevskis, J., Bičevska, Z., & Odītis, I. (2020). User-Oriented Approach to Data Quality Evaluation. *Journal of Universal Computer Science*, *26*(1), 107–126. Retrieved 2023-11-14, from https://lib.jucs.org/article/23992/ doi: 10.3897/jucs.2020.007

Novak, M., Joy, M., & Kermek, D. (2019, May). Source-code Similarity Detection and Detection Tools Used in Academia: A Systematic Review. *ACM Transactions on Computing Education*, *19*(3), 1–37. doi: 10.1145/3313290

Nunamaker, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. *Management Information Systems*, *7*(3), 89–106. doi: 10.1080/07421222.1990.11517898

Ochoa, X., & Duval, E. (2009, August). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, *10*(2), 67–91. doi: 10.1007/s00799-009-0054-4

OECD, EU, & JCR. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide.* OECD Publishing. Retrieved 2024-01-03, from https://www.oecd-ilibrary.org/economics/handbook-on-constructing-composite-indicators-methodology-and-user-guide_9789264043466-en

Offermann, P., Blom, S., Schönherr, M., & Bub, U. (2010). Artifact Types in Information Systems Design Science – A Literature Review. In R. Winter, J. L. Zhao, & S. Aier (Eds.), *Global Perspectives on Design Science Research* (pp. 77–92). St. Gallen, Switzerland: Springer. doi: 10.1007/978-3-642-13335-0_6

Ooms, J. (2023, December). *jsonlite: A Simple and Robust JSON Parser and Generator for R.* Retrieved 2024-01-05, from https://cran.r-project.org/web/packages/jsonlite/index.html

Open Knowledge Foundation. (n.d.). *Why open data?* Retrieved 2024-01-01, from https://okfn.org/en/library/why-open-data/

Palaric, E., Thijs, N., Hammerschmid, G., & Directorate-General for Employment, Social Affairs and Inclusion of the EC. (2018). *A comparative overview of public administration characteristics and performance in EU28* (Tech. Rep. No. KE-02-18-323-EN-N). Luxembourg: Publications Office of the European Union. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/3e89d981-48fc-11e8-be1d-01aa75ed71a1/language-en doi: 10.2767/13319

Palavitsinis, N. (2014). *Metadata quality issues in learning repositories* (PhD dissertation, Universidad de Alcalá). Retrieved 2024-01-03, from https://dialnet.unirioja.es/servlet/tesis?codigo=100796

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017, March). On the Reuse of Scientific Data. *Data Science Journal*, *16*(8), 1–9. Retrieved 2024-01-04, from https://escholarship.org/uc/item/4xf018wx doi: 10.5334/dsj-2017-008

Peffers, K., Tuunanen, T., & Niehaves, B. (2018, March). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research. *European Journal of Information Systems*, *27*(2), 129–139. doi: 10.1080/0960085X.2018.1458066

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007, December). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. doi: 10.2753/MIS0742-1222240302

Permanyer, I. (2011). Assessing the Robustness of Composite Indices Rankings. *Review of Income and Wealth*, *57*(2), 306–326. doi: 10.1111/j.1475-4991.2011.00442.x

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, Massachusetts: Blackwell Publishing.

Pomerantz, J. (2015). *Metadata* (Illustrated edition ed.). England: The MIT Press.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2007, June). *DCMI Abstract Model.* Dublin Core Metadata Initiative. Retrieved 2024-01-05, from https://www.dublincore.org/specifications/dublin-core/abstract-model/

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015, July). A Taxonomy of Evaluation Methods for Information Systems Artifacts. *Journal of Management Information Systems*, *32*(3), 229–267. doi: 10.1080/07421222.2015.1099390

Publications Office of the European Union. (2020, January). *The benefits and value of open data.* Retrieved 2024-01-01, from `https://data.europa.eu/en/publications/datastories/benefits-and-value-open-data`

Publications Office of the European Union. (2021). *Data.europa.eu: Data Quality Guidelines* (Tech. Rep. No. OA-09-21-196-EN-N). Luxembourg: Publications Office of the European Union. doi: 10.2830/879764

Quarati, A. (2021). Open Government Data: Usage trends and metadata quality. *Journal of Information Science*, *49*(4), 887–910. doi: 10.1177/01655515211027775

Reiche, K. J., & Höfig, E. (2013, July). Implementation of Metadata Quality Metrics and Application on Public Government Data. In *Proceedings of the 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops* (pp. 236–241). Retrieved 2023-12-26, from `https://ieeexplore.ieee.org/document/6605795` doi: 10.1109/COMPSACW.2013.32

Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and visualization of metadata quality for government data. In *Proceedings on the International Conference for E-Democracy and Open Governement 2014* (pp. 335–348). Danube University Krems, Austria: Edition Donau-Universität Krems. Retrieved 2023-12-26, from `https://publica.fraunhofer.de/handle/publica/384935`

Republic of Croatia Ministry of Justice and Public Administration. (2015, June). *Strategija razvoja javne uprave za razdoblje od 2015. do 2020. godine.* Hrvatski sabor. Retrieved 2024-01-19, from `https://narodne-novine.nn.hr/clanci/sluzbeni/2015_06_70_1329.html`

Revelle, W. (2023, December). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/psych/index.html`

Riley, J. (2017). *Understanding Metadata: What is metadata, and what is it for?* Baltimore, Maryland: National Information Standards Organization.

Romme, A. (2003). Making a Difference: Organization as Design. *Organization Science*, *14*(5), 558–573. doi: 10.1287/orsc.14.5.558.16769

Ruijer, E. H. J. M., & Martinius, E. (2017, January). Researching the democratic impact of open government data: A systematic literature review. *Information Polity*, *22*(4), 233–250. Retrieved 2024-01-05, from `https://content.iospress.com/articles/information-polity/ip413` doi: 10.3233/IP-170413

Saaty, T. L. (1977, June). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, *15*(3), 234–281. Retrieved 2024-01-05, from `https://www.sciencedirect.com/science/article/pii/0022249677900335` doi: 10.1016/0022-2496(77)90033-5

Saaty, T. L. (1990, September). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26. Retrieved 2024-01-05, from `https://www.sciencedirect.com/science/article/pii/037722179090057I` doi: 10.1016/0377-2217(90)90057-I

Saaty, T. L., & Begičević Ređep, N. (2012, January). The analytic hierarchy process applied to complexity. *International Journal of Economics and Business Research*, *4*(3), 266–283. doi: 10.1504/IJEBR.2012.046821

Saisana, M., Saltelli, A., & Tarantola, S. (2005, March). Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *168*(2), 307–323. doi: 10.1111/j.1467-985X.2005.00350.x

Saltelli, A. (2002, May). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. Retrieved 2024-01-05, from `https://www.sciencedirect.com/science/article/pii/S0010465502002801` doi: 10.1016/S0010-4655(02)00280-1

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., ... Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. John Wiley & Sons.

Schauppenlehner, T., & Muhar, A. (2018, February). Theoretical Availability versus Practical Accessibility: The Critical Role of Metadata Management in Open Data Portals. *Sustainability*, *10*(2), 545. Retrieved 2023-11-13, from `https://www.mdpi.com/2071-1050/10/2/545` doi: 10.3390/su10020545

Signorell, A. (2023, December). *DescTools: Tools for Descriptive Statistics.* Retrieved 2023-12-10, from `https://cran.r-project.org/web/packages/DescTools/index.html`

Simon, H. A. (1996). *The Sciences of the Artificial* (3rd edition ed.). Cambridge, Massachusetts: MIT Press.

Simperl, E., Li, Y., Heath, T., Krieger, B., Hunnius, S., Skoutas, D., & Wenzel, L. (2014). *Open data topologies, catalogues and metadata harmonisation* (Project's work package No. D2.1). Retrieved 2022-11-14, from `https://project.opendatamonitor.eu/wp-content/uploads/deliverable/OpenDataMonitor_611988_D2.1-Open-data-topologies-catalogues-and-metadata-harmonisation.pdf`

Sobol', I. M. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling & Computational Experiment*, *1*, 407–414.

Tauberer, J. (2012). Data Quality: Precision, Accuracy, and Cost. In *Open Government Data.* Retrieved 2024-01-03, from `https://opengovdata.io/2014/data-quality/`

Top 16 Open source Data Portal Solutions for Open Data Publishing. (2019, April). Retrieved 2024-01-04, from `https://medevel.com/15-data-portals-opensource/`

Umbrich, J., Neumaier, S., & Polleres, A. (2015, August). Quality Assessment and Evolution of Open Data Portals. In *2015 3rd International Conference on Future Internet of Things and Cloud* (pp. 404–411). Rome, Italy. Retrieved 2024-01-03, from `https://ieeexplore`

.ieee.org/document/7300846 doi: 10.1109/FiCloud.2015.82

US General Services Administration, US Office of Government and Information Services, & US Office of Management and Budget. (n.d.). *Business case for open data.* Retrieved 2024-01-01, from https://resources.data.gov/resources/open-data/

U.S. Geological Survey. (n.d.). *Archive vs. Repository: Is There a Difference?* Retrieved 2024-01-19, from https://www.usgs.gov/data-management/archive-vs-repository-there-difference

van Aken, J. E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. *Journal of Management Studies*, *41*(2), 219–246. doi: 10.1111/j.1467-6486.2004.00430.x

Van Dooren, W., & Directorate-General for Employment, Social Affairs and Inclusion of the EC. (2018). *Measuring public administration: A feasibility study for better comparative indicators in the EU* (Tech. Rep. No. KE-02-18-989-EN-N). Luxembourg: Publications Office of the European Union. Retrieved from https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=8122&furtherPubs=yes doi: 10.2767/819180

Varga, M., & Vračić, T. (2015). *8. Forum za javnu upravu* (A. Musa, Ed.). Zagreb: Friedrich-Ebert-Stiftung, ured za Hrvatsku, Institut za javnu upravu.

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014, April). Benchmarking open government: An open data perspective. *Government Information Quarterly*, *31*(2), 278–290. Retrieved 2024-01-03, from https://www.sciencedirect.com/science/article/pii/S0740624X14000434 doi: 10.1016/j.giq.2013.10.011

Verhulst, S. G., & Young, A. (2017, July). *OPEN DATA IN DEVELOPING ECONOMIES: Toward Building an Evidence Base on What Works and How* (Tech. Rep.). Governance Lab (GovLab). Retrieved 2024-01-01, from http://odimpact.org

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016, April). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337. Retrieved 2024-01-03, from https://www.sciencedirect.com/science/article/pii/S0740624X16300132 doi: 10.1016/j.giq.2016.02.001

W3C. (2020). *Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation 04 February 2020.* W3C. Retrieved from https://www.w3.org/TR/vocab-dcat-2/

W3C. (2023). *Data Catalog Vocabulary (DCAT) - Version 3. W3C Working Draft 07 March 2023.* W3C. Retrieved from https://www.w3.org/TR/vocab-dcat-3/

Wand, Y., & Wang, R. Y. (1996, November). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. doi: 10.1145/240455.240479

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2023, February). Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy*. doi: 10.1007/s13132-022-01096-6

Wang, V., & Shepherd, D. (2020, January). Exploring the extent of openness of open government data – A critique of open government datasets in the UK. *Government Information Quarterly*, *37*(1). doi: 10.1016/j.giq.2019.101405

Watson, R. T. (2016). *Data Management: Databases and Organizations* (6th edition ed.). Prospect Press.

Wei, T., & Simko, V. (2021, November). *corrplot: Visualization of a Correlation Matrix.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/corrplot/index.html`

Wickham, H. (2023a, August). *httr: Tools for Working with URLs and HTTP.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/httr/index.html`

Wickham, H. (2023b, November). *stringr: Simple, Consistent Wrappers for Common String Operations.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/stringr/index.html`

Wickham, H. (2023c, February). *tidyverse: Easily Install and Load the 'Tidyverse'.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/tidyverse/index.html`

Wickham, H., & Bryan, J. (2023, July). *readxl: Read Excel Files.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/readxl/index.html`

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., ... Dunnington, D. (2023, October). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* Retrieved 2024-01-09, from `https://cran.r-project.org/web/packages/ggplot2/index.html`

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023, November). *dplyr: A Grammar of Data Manipulation.* Retrieved 2024-01-09, from `https://cran.r-project.org/web/packages/dplyr/index.html`

Wong, T.-T. (1998, December). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, *97*(2), 165–181. doi: 10.1016/S0096-3003(97)10140-0

Yang, J., Wittern, E., Ying, A. T. T., Dolby, J., & Tan, L. (2018, May). Towards extracting web API specifications from documentation. In *Proceedings of the 15th International Conference on Mining Software Repositories* (pp. 454–464). Gothenburg, Sweden: Association for Computing Machinery. doi: 10.1145/3196398.3196411

Young, A., Zahuranec, A. J., Verhulst, S. G., & Gazaryan, K. (2021, March). *The Third Wave of Open Data Toolkit: Operational Guidance on Capturing the Institutional and Societal Value of Data Re-Use* (Tech. Rep.). New York: New York University Tandon School of Engineering.

Zabala, A., Held, M., & Hermans, F. (2023, March). *qmethod: Analysis of Subjective Perspectives Using Q Methodology.* Retrieved 2024-01-05, from `https://cran.r-project.org/web/packages/qmethod/index.html`

Zeng, M. L., & Qin, J. (2022). *Metadata* (Third edition ed.). Chicago: ALA Neal-Schuman.

Zhang, H., & Xiao, J. (2020). Quality assessment framework for open government data: Meta-synthesis of qualitative research, 2009-2019. *The Electronic Library*, *38*(2), 209–222. doi: 10.1108/EL-06-2019-0145

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in Estimation and Hypothesis Testing of Correlation. *Psicológica*, *24*(1), 133–158.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Sheikh Alibaks, R. (2012, December). Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, *10*(2), 156-172. Retrieved 2024-01-03, from `https://academic-publishing.org/index.php/ejeg/article/view/571`

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., ... Sinz, E. J. (2010). Memorandum on design-oriented information systems research. *European Journal of Information Systems*, *20*(1), 7–10. doi: 10.1057/ejis.2010.55

Šlibar, B. (2019). Framework for Quality Assessment of Open Datasets. In *Proceedings of the 11th International Doctoral Seminar (IDS 2019)* (pp. 19–21). Varaždin, Croatia: Faculty of Organization and Informatics and Faculty of Materials Science and Technology in Trnava. Retrieved 2024-01-04, from `https://www.bib.irb.hr:8443/1053442`

Šlibar, B. (2024a, January). *Demonstration of the composite indicator of metadata quality for open datasets.* Harvard Dataverse. (Available upon request) doi: 10.7910/DVN/VF2Q9F

Šlibar, B. (2024b, January). *Harmonisation of metadata statements from different metadata standards.* Harvard Dataverse. (Available upon request) doi: 10.7910/DVN/M3MINM

Šlibar, B. (2024c, January). *Theoretical framework of metadata quality for open datasets – initial version.* Harvard Dataverse. (Available upon request) doi: 10.7910/DVN/COW8EI

Šlibar, B., & Mu, E. (2022, October). OGD metadata country portal publishing guidelines compliance: A multi-case study search for completeness and consistency. *Government Information Quarterly*, *39*(4). Retrieved 2024-01-03, from `https://www.sciencedirect.com/science/article/pii/S0740624X22000922` doi: 10.1016/j.giq.2022.101756

Šlibar, B., Oreški, D., & Begičević Ređep, N. (2021, April). Importance of the Open Data Assessment: An Insight Into the (Meta) Data Quality Dimensions. *SAGE Open*, *11*(2), 1–18. doi: 10.1177/21582440211023178

Šlibar, B., Oreški, D., & Kliček, B. (2018). Aspects of open data and illustrative quality metrics: literature review. In *Economic and Social Development 35th International Scientific Conference on Economic and Social Development – "Sustainability from an Economic and Social Perspective" - Book of Proceedings* (pp. 90–99). Retrieved 2024-01-03, from `https://www.bib.irb.hr:8443/968633`

APPENDIX A

INVITATION TO PARTICIPATE IN THE TESTING CONTENT VALIDITY OF THE INITIAL THEORETICAL FRAMEWORK

Greetings,

I invite You to participate as an expert in the research related to my dissertation *Quality assessment of open datasets metadata* at the doctoral study in Information Sciences at the University of Zagreb, Faculty of Organization and Informatics. I have collected Your contact or/and Your consent to participate in the research at one of the international scientific or professional conferences (e.g., EU DataViz 2021), summer schools (e.g., Summer School of the Twinning Open Data Operational project), or some other event.

Short summary of the dissertation: Open data are an extremely valuable information technology resource for economic, social, and human development that adds new values to society's development. More and more countries in the world are establishing portals at the national, regional, and local levels, and thus the amount of available open data is growing. The usability of open data depends on the quality of their metadata, whose evaluation is an open research question. The objective of the proposed research is to develop a theoretical framework of open metadata quality and operationalise it through a new composite indicator that will enable the comparison of open datasets metadata. The research approach will be based on the Methodological framework for design science research and Methodology for constructing composite indicators. The scientific contribution will be achieved through the development of a framework and composite indicator, a better understanding of the concept of open (meta)data quality, and empirical research of the public sector metadata quality.

Purpose of Your Participation: Testing content validity of the proposed theoretical framework.

If You decide to participate, Your engagement includes the following:

- Your opinion on which individual indicators are relevant for a particular quality dimension of open data metadata will be collected by the EQ Web Sort tool. You will perform Q sorting task for each quality dimension separately. Once You start a Q sorting task for some dimension, complete it. Otherwise, You will have to start over again. You can perform Q sorting for different dimensions at different times. Instructions on how to sort the individual indicators will be provided within the tool.

- The individual indicator represents a base to make an assessment in relation to a given objective. Here, an individual indicator refers to a combination of metrics (e.g., Existence) and a metadata field (e.g., File format). The core set of metadata fields is proposed based on ISO/IEC 11179 Information technology — Metadata registries (ISO/IEC 11179) and Data Catalog Vocabulary - Version 2 (DCAT 2). Therefore, the tag for metadata in the individual indicator consists of the class name and attribute according to ISO/IEC 11179 and DCAT 2. For example, instead of the tag File format, there will be Data_Set_Distribution ->format; Distribution ->format. You can find mappings of fields and their description in the 'Descriptions of metadata fields' document.

- The theoretical framework is proposed based on the literature review. It consists of the following five dimensions Findability ('F'), Retrievability ('R'), Interoperability ('I'), Reusability ('U'), and Contextuality ('C') at the highest level. A description of each dimension is in the 'Definitions of dimensions' document. Please read it carefully. Also, additional literature is attached to this email.

- Q sorts contain a different number of statements and the time required to complete each of them may vary. However, the most complex one shouldn't take more than an hour to complete. **Please complete all Q sorts by Tuesday, 3 January 2023 at the latest.**

  Link to access Q sorts: https://services.foi.hr/qsort/ You will need a code for the sign-in process. Your participant code is: [Participant ID]

- Your participation in this research is entirely voluntary. Please note that You don't have to take part in this research if You don't wish to do so. Also, You can withdraw from the participation at any time without explaining the reason for doing so. However, I would kindly ask You to complete the Q sorting task for all dimensions if You have done it for one.

If You want to find out more about me check https://www.foi.unizg.hr/en/staff/barbara.slibar I sincerely hope You will find some time to help me with my research.

Thank You in advance,

Barbara Šlibar

APPENDIX B

INVITATION TO PARTICIPATE IN THE WEIGHTING OF THE ELEMENTS OF THE THEORETICAL FRAMEWORK

Dear Mr/Ms [Recipient surname],

Thank You once more for participating in testing content validity of the proposed theoretical framework in the research related to my dissertation *Quality assessment of open datasets metadata* at the doctoral study in Information Sciences at the University of Zagreb Faculty of Organization and Informatics. I only need one additional input from You for my research.

If You decide to participate, please fill out the **survey in the excel document 'The importance of elements in the theoretical framework'**. The expected time to complete survey is 45 minutes. You can find instructions on how to fill out the survey in the following video `https://youtu.be/IclCMKBeM38`. **Please complete the survey by June 11th, 2023 at the latest.**

Purpose of Your Participation: Weighting of the elements of the theoretical framework.

One of the objectives of my research is to operationalise developed theoretical framework of open metadata quality through a new composite indicator that will enable the comparison of open datasets metadata. Since research approach is based on the Methodology for constructing composite indicators according to which individual indicators should be weighted and aggregated with respect to the developed theoretical framework, I would like to have your opinion on relative importance of elements of the developed theoretical framework.

Additional information:

- The theoretical framework consists of the following five dimensions Findability ('F'), Retrievability ('R'), Interoperability ('I'), Reusability ('U'), and Contextuality ('C') at the highest level. A description of each dimension is in the 'Definitions of dimensions' document. The individual indicator represents a base to make an assessment in relation to a given objective. Here, an individual indicator refers to a combination of metrics (e.g., Existence) and a metadata field (e.g., File format). The core set of metadata fields is proposed based on ISO/IEC 11179 Information technology — Metadata registries (ISO/IEC 11179) and Data Catalog Vocabulary - Version 2 (DCAT 2). Therefore, the tag for metadata in the individual indicator consists of the class name and attribute according to ISO/IEC 11179 and DCAT 2. For example, instead of the tag File format, there will be Data_Set_Distribution ->format; Distribution ->format. You can find mappings of fields and their description in the 'Descriptions of metadata fields' document.

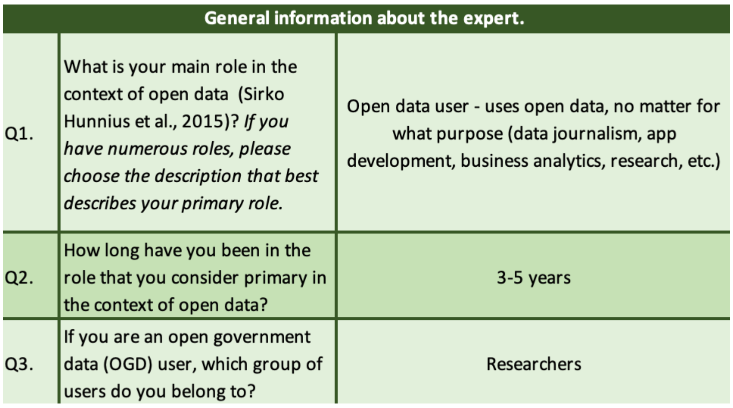
- Your participation in this research is entirely voluntary. Please note that You don’t have to take part in this research if You don’t wish to do so. Also, You can withdraw from the participation at any time without explaining the reason for doing so.

I sincerely look forward to Your opinion,

Barbara Šlibar

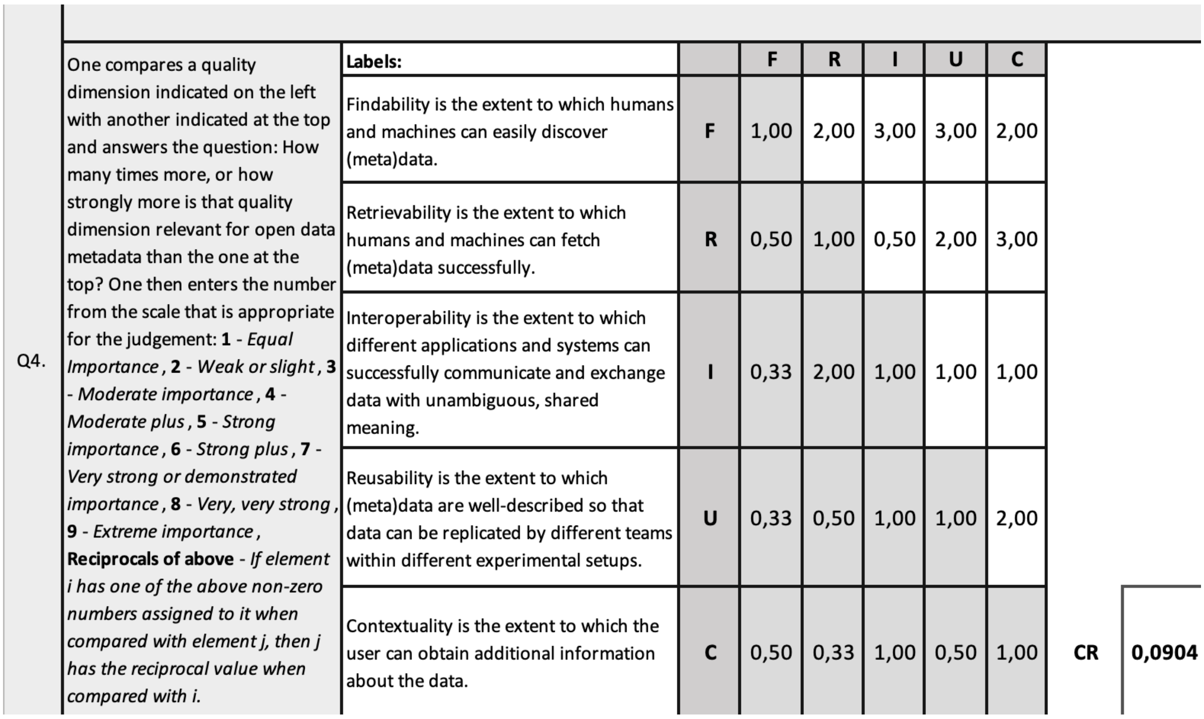
APPENDIX C

SURVEY ON RELATIVE IMPORTANCE OF ELEMENTS IN THE THEORETICAL FRAMEWORK



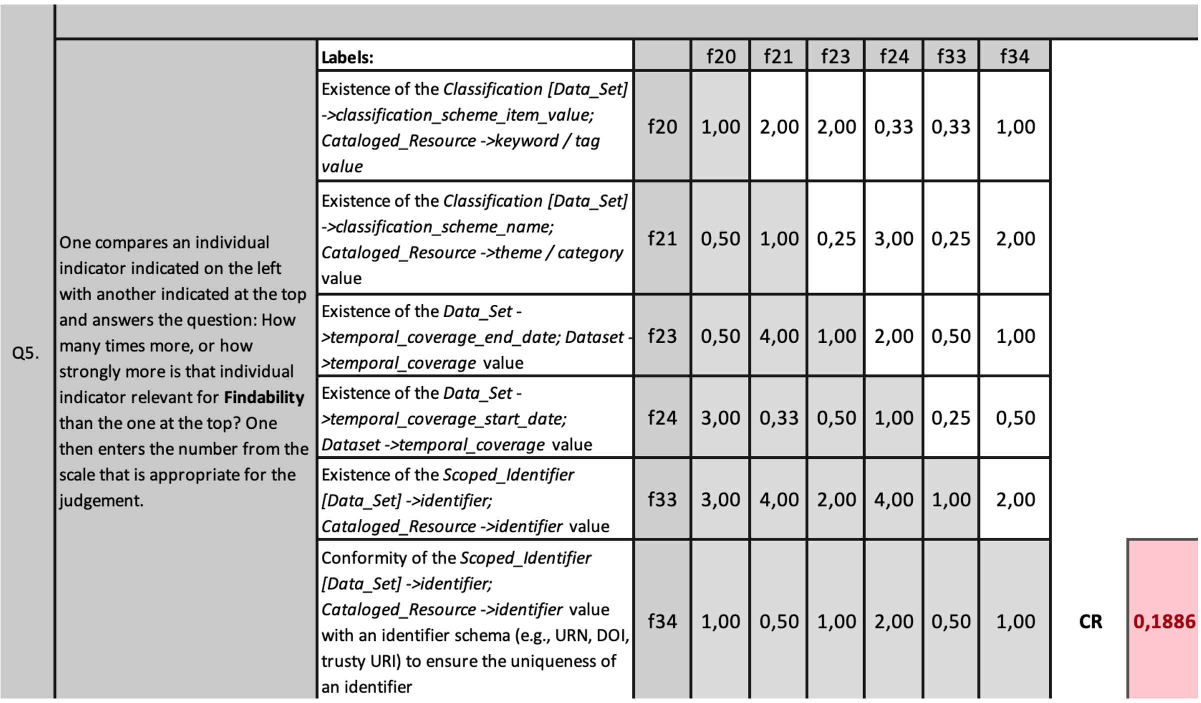
General information about the expert.		
Q1.	What is your main role in the context of open data (Sirko Hunnius et al., 2015)? *If you have numerous roles, please choose the description that best describes your primary role.*	Open data user - uses open data, no matter for what purpose (data journalism, app development, business analytics, research, etc.)
Q2.	How long have you been in the role that you consider primary in the context of open data?	3-5 years
Q3.	If you are an open government data (OGD) user, which group of users do you belong to?	Researchers

Figure C.1: Survey section for the general information about the expert.



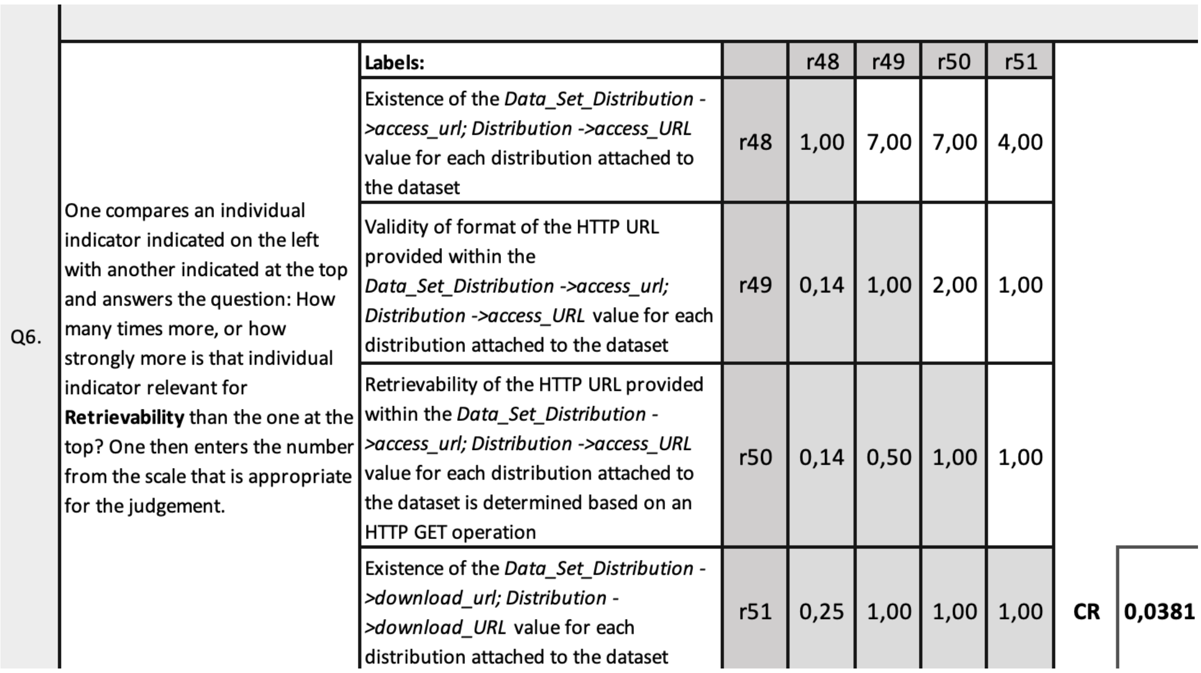
		Labels:		**F**	**R**	**I**	**U**	**C**		
Q4.	One compares a quality dimension indicated on the left with another indicated at the top and answers the question: How many times more, or how strongly more is that quality dimension relevant for open data metadata than the one at the top? One then enters the number from the scale that is appropriate for the judgement: **1** - *Equal Importance* , **2** - *Weak or slight* , **3** - *Moderate importance* , **4** - *Moderate plus* , **5** - *Strong importance* , **6** - *Strong plus* , **7** - *Very strong or demonstrated importance* , **8** - *Very, very strong* , **9** - *Extreme importance* , **Reciprocals of above** - *If element i has one of the above non-zero numbers assigned to it when compared with element j, then j has the reciprocal value when compared with i.*	Findability is the extent to which humans and machines can easily discover (meta)data.	**F**	1,00	2,00	3,00	3,00	2,00		
		Retrievability is the extent to which humans and machines can fetch (meta)data successfully.	**R**	0,50	1,00	0,50	2,00	3,00		
		Interoperability is the extent to which different applications and systems can successfully communicate and exchange data with unambiguous, shared meaning.	**I**	0,33	2,00	1,00	1,00	1,00		
		Reusability is the extent to which (meta)data are well-described so that data can be replicated by different teams within different experimental setups.	**U**	0,33	0,50	1,00	1,00	2,00		
		Contextuality is the extent to which the user can obtain additional information about the data.	**C**	0,50	0,33	1,00	0,50	1,00	**CR**	**0,0904**

Figure C.2: Survey section for the pairwise comparisons of the dimensions with respect to the composite indicator (Legend: CR - consistency ratio).



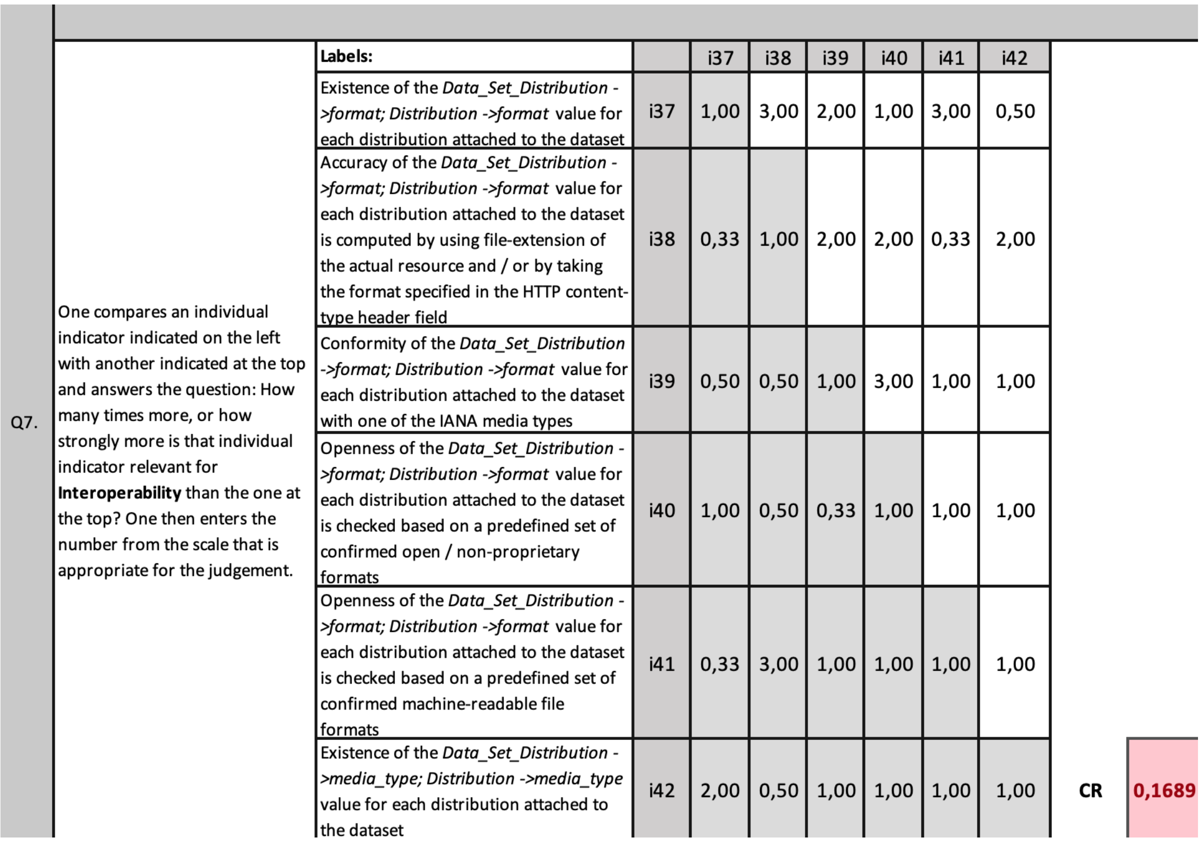
		Labels:		f20	f21	f23	f24	f33	f34		
Q5.	One compares an individual indicator indicated on the left with another indicated at the top and answers the question: How many times more, or how strongly more is that individual indicator relevant for **Findability** than the one at the top? One then enters the number from the scale that is appropriate for the judgement.	Existence of the *Classification [Data_Set] ->classification_scheme_item_value; Cataloged_Resource ->keyword / tag value*	f20	1,00	2,00	2,00	0,33	0,33	1,00		
		Existence of the *Classification [Data_Set] ->classification_scheme_name; Cataloged_Resource ->theme / category* value	f21	0,50	1,00	0,25	3,00	0,25	2,00		
		Existence of the *Data_Set ->temporal_coverage_end_date; Dataset ->temporal_coverage* value	f23	0,50	4,00	1,00	2,00	0,50	1,00		
		Existence of the *Data_Set ->temporal_coverage_start_date; Dataset ->temporal_coverage* value	f24	3,00	0,33	0,50	1,00	0,25	0,50		
		Existence of the *Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier* value	f33	3,00	4,00	2,00	4,00	1,00	2,00		
		Conformity of the *Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier* value with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier	f34	1,00	0,50	1,00	2,00	0,50	1,00	**CR**	**0,1886**

Figure C.3: Survey section for the pairwise comparisons of the individual indicators with respect to the findability dimension (Legend: CR - consistency ratio).



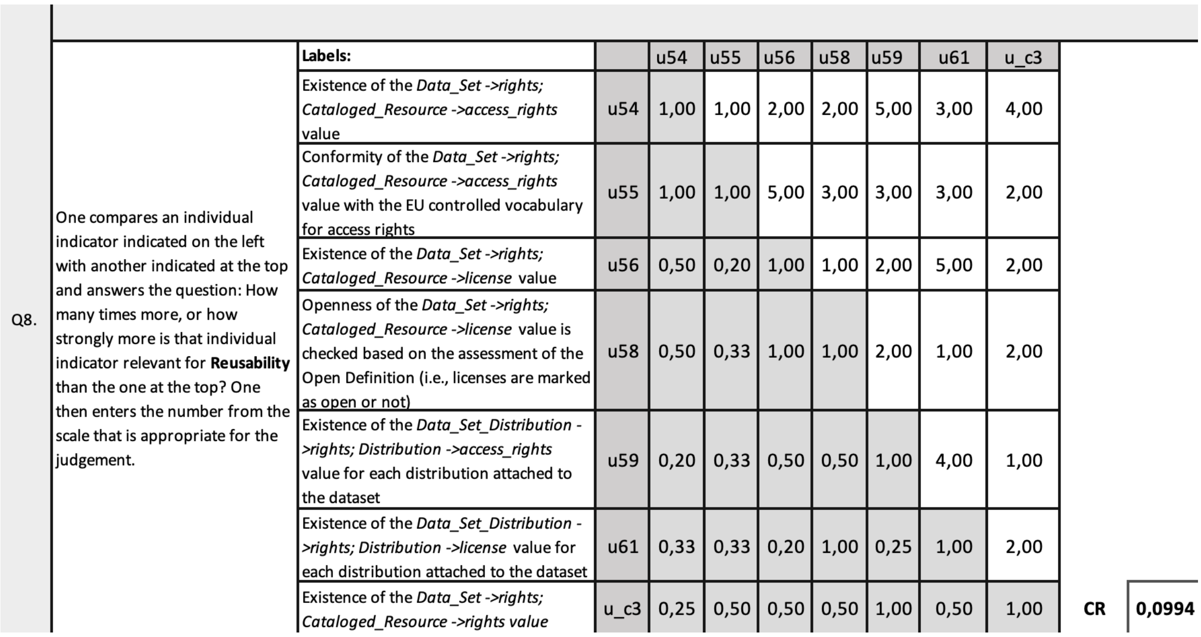
		Labels:		r48	r49	r50	r51		
Q6.	One compares an individual indicator indicated on the left with another indicated at the top and answers the question: How many times more, or how strongly more is that individual indicator relevant for **Retrievability** than the one at the top? One then enters the number from the scale that is appropriate for the judgement.	Existence of the *Data_Set_Distribution ->access_url; Distribution ->access_URL* value for each distribution attached to the dataset	r48	1,00	7,00	7,00	4,00		
		Validity of format of the HTTP URL provided within the *Data_Set_Distribution ->access_url; Distribution ->access_URL* value for each distribution attached to the dataset	r49	0,14	1,00	2,00	1,00		
		Retrievability of the HTTP URL provided within the *Data_Set_Distribution ->access_url; Distribution ->access_URL* value for each distribution attached to the dataset is determined based on an HTTP GET operation	r50	0,14	0,50	1,00	1,00		
		Existence of the *Data_Set_Distribution ->download_url; Distribution ->download_URL* value for each distribution attached to the dataset	r51	0,25	1,00	1,00	1,00	**CR**	**0,0381**

Figure C.4: Survey section for the pairwise comparisons of the individual indicators with respect to the retrievability dimension (Legend: CR - consistency ratio).



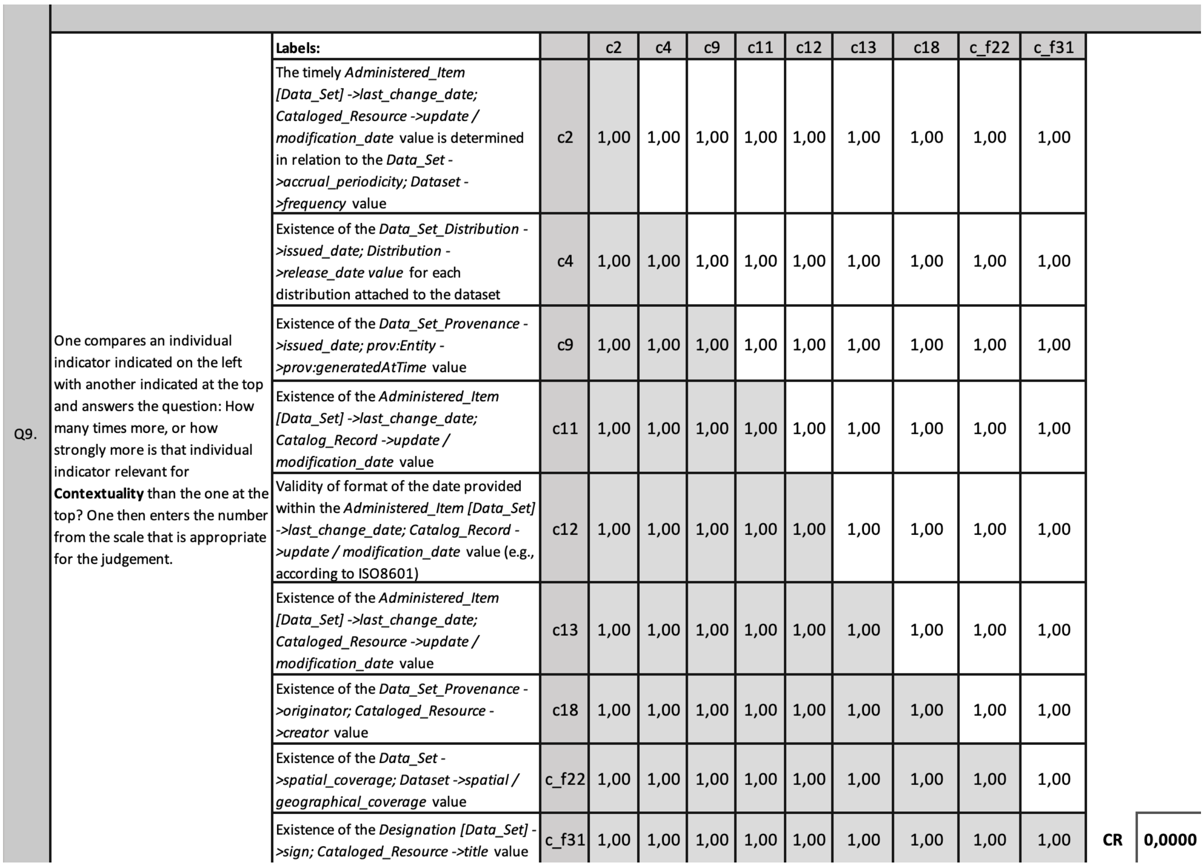
		Labels:		i37	i38	i39	i40	i41	i42		
Q7.	One compares an individual indicator indicated on the left with another indicated at the top and answers the question: How many times more, or how strongly more is that individual indicator relevant for **Interoperability** than the one at the top? One then enters the number from the scale that is appropriate for the judgement.	Existence of the *Data_Set_Distribution ->format; Distribution ->format* value for each distribution attached to the dataset	i37	1,00	3,00	2,00	1,00	3,00	0,50		
		Accuracy of the *Data_Set_Distribution ->format; Distribution ->format* value for each distribution attached to the dataset is computed by using file-extension of the actual resource and / or by taking the format specified in the HTTP content-type header field	i38	0,33	1,00	2,00	2,00	0,33	2,00		
		Conformity of the *Data_Set_Distribution ->format; Distribution ->format* value for each distribution attached to the dataset with one of the IANA media types	i39	0,50	0,50	1,00	3,00	1,00	1,00		
		Openness of the *Data_Set_Distribution ->format; Distribution ->format* value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	i40	1,00	0,50	0,33	1,00	1,00	1,00		
		Openness of the *Data_Set_Distribution ->format; Distribution ->format* value for each distribution attached to the dataset is checked based on a predefined set of confirmed machine-readable file formats	i41	0,33	3,00	1,00	1,00	1,00	1,00		
		Existence of the *Data_Set_Distribution ->media_type; Distribution ->media_type* value for each distribution attached to the dataset	i42	2,00	0,50	1,00	1,00	1,00	1,00	CR	0,1689

Figure C.5: Survey section for the pairwise comparisons of the individual indicators with respect to the interoperability dimension (Legend: CR - consistency ratio).



		Labels:		u54	u55	u56	u58	u59	u61	u_c3		
Q8.	One compares an individual indicator indicated on the left with another indicated at the top and answers the question: How many times more, or how strongly more is that individual indicator relevant for **Reusability** than the one at the top? One then enters the number from the scale that is appropriate for the judgement.	Existence of the *Data_Set ->rights; Cataloged_Resource ->access_rights* value	u54	1,00	1,00	2,00	2,00	5,00	3,00	4,00		
		Conformity of the *Data_Set ->rights; Cataloged_Resource ->access_rights* value with the EU controlled vocabulary for access rights	u55	1,00	1,00	5,00	3,00	3,00	3,00	2,00		
		Existence of the *Data_Set ->rights; Cataloged_Resource ->license* value	u56	0,50	0,20	1,00	1,00	2,00	5,00	2,00		
		Openness of the *Data_Set ->rights; Cataloged_Resource ->license* value is checked based on the assessment of the Open Definition (i.e., licenses are marked as open or not)	u58	0,50	0,33	1,00	1,00	2,00	1,00	2,00		
		Existence of the *Data_Set_Distribution ->rights; Distribution ->access_rights* value for each distribution attached to the dataset	u59	0,20	0,33	0,50	0,50	1,00	4,00	1,00		
		Existence of the *Data_Set_Distribution ->rights; Distribution ->license* value for each distribution attached to the dataset	u61	0,33	0,33	0,20	1,00	0,25	1,00	2,00		
		Existence of the *Data_Set ->rights; Cataloged_Resource ->rights value*	u_c3	0,25	0,50	0,50	0,50	1,00	0,50	1,00	CR	0,0994

Figure C.6: Survey section for the pairwise comparisons of the individual indicators with respect to the reusability dimension (Legend: CR - consistency ratio).



		Labels:		c2	c4	c9	c11	c12	c13	c18	c_f22	c_f31		
Q9.	One compares an individual indicator indicated on the left with another indicated at the top and answers the question: How many times more, or how strongly more is that individual indicator relevant for **Contextuality** than the one at the top? One then enters the number from the scale that is appropriate for the judgement.	The timely *Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* value is determined in relation to the *Data_Set ->accrual_periodicity; Dataset ->frequency* value	c2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Data_Set_Distribution ->issued_date; Distribution ->release_date value* for each distribution attached to the dataset	c4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime* value	c9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date* value	c11	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Validity of format of the date provided within the *Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date* value (e.g., according to ISO8601)	c12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date* value	c13	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Data_Set_Provenance ->originator; Cataloged_Resource ->creator* value	c18	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Data_Set ->spatial_coverage; Dataset ->spatial / geographical_coverage* value	c_f22	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00		
		Existence of the *Designation [Data_Set] ->sign; Cataloged_Resource ->title* value	c_f31	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	**CR**	**0,0000**

Figure C.7: Survey section for the pairwise comparisons of the individual indicators with respect to the contextuality dimension (Legend: CR - consistency ratio).

APPENDIX D

AVERAGE SCORE AND LAWSHE'S CONTENT VALIDITY RATIO OF THE INDIVIDUAL INDICATORS OF THE INITIAL THEORETICAL FRAMEWORK

Table D.1: Mean Q sort rank ($\bar{x}$) and Lawshe's content validity ratio (CVR) together with the corresponding p-values based on the t-test and the exact binomial test (BT) for the individual indicators of the findability dimension in the initial version of the theoretical framework.

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f20	Existence of the Classification [Data_Set] ->classification_scheme_item_value; Cataloged_Resource ->keyword / tag value	0.545	0.400	0.001	0.818	0.033
f21	Existence of the Classification [Data_Set] ->classification_scheme_name; Cataloged_Resource ->theme / category value	0.364	0.424	0.009	0.636	0.127
f33	Existence of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value	0.341	0.375	0.007	0.636	0.127
f31	Existence of the Designation [Data_Set] ->sign; Cataloged_Resource ->title value	0.273	0.617	0.087	0.455	0.311
f34	Conformity of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier	0.250	0.418	0.038	0.818	0.033
c18	Existence of the Data_Set_Provenance ->originator; Cataloged_Resource ->creator value	0.250	0.461	0.051	0.455	0.311
f22	Existence of the Data_Set ->spatial_coverage; Dataset ->spatial / geographical_coverage value	0.227	0.361	0.032	0.636	0.127
f24	Existence of the Data_Set ->temporal_coverage_start_date; Dataset ->temporal_coverage value	0.205	0.368	0.047	0.636	0.127
i37	Existence of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset	0.205	0.313	0.028	0.818	0.033

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f23	Existence of the Data_Set ->temporal_coverage_end_date; Dataset ->temporal_coverage value	0.182	0.372	0.068	0.636	0.127
i40	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	0.159	0.257	0.033	0.818	0.033
f25	Existence of the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value	0.136	0.438	0.163	0.273	0.549
i42	Existence of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset	0.114	0.452	0.212	0.273	0.549
i38	Accuracy of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is computed by using file-extension of the actual resource and / or by taking the format specified in the HTTP content-type header field	0.045	0.416	0.362	0.273	0.549
u71	Existence of the Data_Set_Provenance ->generation_type; Dataset ->was_generated_by value	0.045	0.270	0.294	0.455	0.311
i45	Openness of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	0.023	0.261	0.389	0.455	0.311
c13	Existence of the Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value	0.023	0.395	0.426	0.455	0.311

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
u66	Validity of format of the HTTP URL provided within Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset	-0.023	0.518	0.556	0.273	0.549
c10	Validity of format of the date provided within Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime value (e.g., according to ISO 8601)	-0.045	0.292	0.692	0.091	0.767
c15	Existence of the N/A; Distribution ->update / modification_date value for each distribution attached to the dataset	-0.068	0.298	0.767	-0.091	0.908
u60	Conformity of the Data_Set_Distribution ->rights; Distribution ->access_rights value for each distribution attached to the dataset with the EU controlled vocabulary for access rights	-0.068	0.276	0.784	0.091	0.767
i43	Accuracy of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-type header field	-0.091	0.322	0.815	0.091	0.767
c16	Validity of format of the date provided within N/A; Distribution ->update / modification_date value for each distribution attached to the dataset (e.g., according to ISO 8601)	-0.091	0.491	0.724	0.091	0.767
f36	Conformity of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier	-0.114	0.540	0.749	0.273	0.549

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f35	Existence of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset	-0.136	0.342	0.892	0.091	0.767
f32	Semantic distance between the Designation [Data_Set] ->sign; Cataloged_Resource ->title value AND the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value	-0.159	0.340	0.924	-0.091	0.908
f30	Semantic distance between the Designation [Data_Set_Distribution] ->sign; Distribution ->title value for each distribution attached to the dataset AND the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset	-0.205	0.400	0.940	-0.273	0.973
u68	Validity of format of the email address provided within Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value	-0.318	0.337	0.995	-0.636	0.999
f26	Readability of the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is computed by using the Flesch-Kincaid Reading Ease test	-0.341	0.584	0.959	-0.273	0.973
c8	Accuracy of the Data_Set_Distribution ->size; Distribution ->byteSize value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-length header field	-0.364	0.517	0.979	-0.273	0.973
f28	Readability of the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is computed by using the Flesch-Kincaid Reading Ease test	-0.386	0.377	0.997	-0.636	0.999
f29	Intrinsic precision of the text provided within Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is determined by spelling mistakes	-0.409	0.465	0.992	-0.455	0.995

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f27	Intrinsic precision of the text provided within Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is determined by spelling mistakes	-0.568	0.298	1.000	-0.818	1.000

Table D.2: Mean Q sort rank ($\bar{x}$) and Lawshe's content validity ratio (CVR) together with the corresponding p-values based on the t-test and the exact binomial test (BT) for the individual indicators of the retrievability dimension in the initial version of the theoretical framework.

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
r48	Existence of the Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset	0.667	0.354	0.000	1.000	0.026
r50	Retrievability of the HTTP URL provided within Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset is determined based on an HTTP GET operation	0.500	0.433	0.004	1.000	0.026
r49	Validity of format of the HTTP URL provided within Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset	0.389	0.220	0.000	1.000	0.026
r51	Existence of the Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset	0.333	0.433	0.025	1.000	0.026
r52	Validity of format of the HTTP URL provided within Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset	0.000	0.354	0.500	0.556	0.377
r53	Retrievability of the HTTP URL provided within Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset is determined based on an HTTP GET operation	0.000	0.354	0.500	0.556	0.377

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f33	Existence of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value	-0.056	0.527	0.620	0.333	0.650
f36	Conformity of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier	-0.167	0.354	0.902	0.111	0.855
c16	Validity of format of the date provided within N/A; Distribution ->update / modification_date value for each distribution attached to the dataset (e.g., according to ISO 8601)	-0.333	0.500	0.960	-0.111	0.958
u56	Existence of the Data_Set ->rights; Cataloged_Resource ->license value	-0.389	0.486	0.978	-0.333	0.992
c8	Accuracy of the Data_Set_Distribution ->size; Distribution ->byteSize value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-length header field	-0.444	0.527	0.982	-0.333	0.992
u57	Conformity of the Data_Set ->rights; Cataloged_Resource ->license value with one of the licenses from the predefined list provided by the Open Definition or the EU Vocabularies related to licenses	-0.500	0.250	1.000	-0.778	1.000

Table D.3: Mean Q sort rank ($\bar{x}$) and Lawshe's content validity ratio (CVR) together with the corresponding p-values based on the t-test and the exact binomial test (BT) for the individual indicators of the interoperability dimension in the initial version of the theoretical framework.

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
i41	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed machine-readable file formats	0.370	0.261	0.001	1.000	0.013

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
i40	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	0.296	0.455	0.043	0.556	0.269
i42	Existence of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset	0.296	0.261	0.005	1.000	0.013
r51	Existence of the Data_Set_Distribution ->download_url; Distribution ->download_URL value for each distribution attached to the dataset	0.296	0.564	0.077	0.333	0.531
i37	Existence of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset	0.259	0.364	0.033	0.778	0.087
i39	Conformity of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset with one of the IANA media types	0.222	0.471	0.098	0.556	0.269
u56	Existence of the Data_Set ->rights; Cataloged_Resource ->license value	0.185	0.603	0.192	0.333	0.531
f21	Existence of the Classification [Data_Set] ->classification_scheme_name; Cataloged_Resource ->theme / category value	0.185	0.444	0.123	0.556	0.269
i45	Openness of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	0.148	0.444	0.173	0.333	0.531
i38	Accuracy of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is computed by using file-extension of the actual resource and / or by taking the format specified in the HTTP content-type header field	0.111	0.373	0.199	0.778	0.087

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
i44	Conformity of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset with one of the IANA media types	0.111	0.441	0.236	0.556	0.269
u54	Existence of the Data_Set ->rights; Cataloged_Resource ->access_rights value	-0.037	0.588	0.573	0.111	0.772
i43	Accuracy of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-type header field	-0.074	0.494	0.668	0.333	0.531
u55	Conformity of the Data_Set ->rights; Cataloged_Resource ->access_rights value with the EU controlled vocabulary for access rights	-0.111	0.373	0.801	0.111	0.772
c6	Existence of the Data_Set_Distribution ->rights; Distribution ->rights value for each distribution attached to the dataset	-0.111	0.441	0.764	0.111	0.772
c15	Existence of the N/A; Distribution ->update / modification_date value for each distribution attached to the dataset	-0.148	0.503	0.799	-0.111	0.920
i46	Accuracy of the Registration_Authority [Data_Set] ->documentation_language_identifier; Cataloged_Resource ->language value is computed by using language detection on the actual resource and / or HTTP content-language header field	-0.185	0.580	0.817	-0.111	0.920
i47	Conformity of the Registration_Authority [Data_Set] ->documentation_language_identifier; Cataloged_Resource ->language value to a given standard such as ISO 639	-0.259	0.494	0.923	-0.333	0.981
f20	Existence of the Classification [Data_Set] ->classification_scheme_item_value; Cataloged_Resource ->keyword / tag value	-0.259	0.494	0.923	-0.333	0.981

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f28	Readability of the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is computed by using the Flesch-Kincaid Reading Ease test	-0.593	0.364	0.999	-0.778	1.000
f27	Intrinsic precision of the text provided within Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is determined by spelling mistakes	-0.704	0.309	1.000	-0.778	1.000

Table D.4: Mean Q sort rank ($\bar{x}$) and Lawshe's content validity ratio (CVR) together with the corresponding p-values based on the t-test and the exact binomial test (BT) for the individual indicators of the reusability dimension in the initial version of the theoretical framework.

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
u56	Existence of the Data_Set ->rights; Cataloged_Resource ->license value	0.556	0.391	0.001	0.778	0.071
u58	Openness of the Data_Set ->rights; Cataloged_Resource ->license value is checked based on the assessment of the Open Definition (i.e., licenses are marked as open or not)	0.500	0.500	0.009	0.778	0.071
u54	Existence of the Data_Set ->rights; Cataloged_Resource ->access_rights value	0.472	0.441	0.006	0.556	0.232
u61	Existence of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset	0.444	0.410	0.006	0.778	0.071
u59	Existence of the Data_Set_Distribution ->rights; Distribution ->access_rights value for each distribution attached to the dataset	0.333	0.395	0.018	0.778	0.071
i40	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	0.250	0.375	0.040	0.778	0.071
u55	Conformity of the Data_Set ->rights; Cataloged_Resource ->access_rights value with the EU controlled vocabulary for access rights	0.222	0.458	0.092	0.556	0.232

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
u57	Conformity of the Data_Set ->rights; Cataloged_Resource ->license value with one of the licenses from the predefined list provided by the Open Definition or the EU Vocabularies related to licenses	0.222	0.491	0.106	0.556	0.232
u64	Existence of the Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or the Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset	0.222	0.491	0.106	0.556	0.232
c3	Existence of the Data_Set ->rights; Cataloged_Resource ->rights value	0.194	0.300	0.044	0.778	0.071
u62	Conformity of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset with one of the licenses from the predefined list provided by the Open Definition related to licenses	0.083	0.217	0.141	0.556	0.232
c11	Existence of the Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date value	0.083	0.280	0.199	0.778	0.071
c15	Existence of the N/A; Distribution ->update / modification_date value for each distribution attached to the dataset	0.056	0.300	0.297	0.333	0.483
c4	Existence of the Data_Set_Distribution ->issued_date; Distribution ->release_date value for each distribution attached to the dataset	0.056	0.349	0.323	0.556	0.232
u63	Openness of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset is checked based on the assessment of the Open Definition (i.e., licenses are marked as open or not)	0.028	0.317	0.400	0.333	0.483

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
c17	Existence of the Registration_State [Data_Set] ->effective_date; Cataloged_Resource ->release_date value	0.028	0.441	0.427	0.111	0.733
u60	Conformity of the Data_Set_Distribution ->rights; Distribution ->access_rights value for each distribution attached to the dataset with the EU controlled vocabulary for access rights	0.000	0.484	0.500	0.333	0.483
f20	Existence of the Classification [Data_Set] ->classification_scheme_item_value; Cataloged_Resource ->keyword / tag value	0.000	0.515	0.500	0.333	0.483
u66	Validity of format of the HTTP URL provided within Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset	-0.028	0.507	0.563	0.333	0.483
c18	Existence of the Data_Set_Provenance ->originator; Cataloged_Resource ->creator value	-0.028	0.317	0.600	0.556	0.232
u65	Validity of format of the email address provided within Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value assigned to dataset or Data_Set_Distribution | Submission_Record [Data_Set / Data_Set_Distribution] ->distributor | organization, contact; Cataloged_Resource ->publisher value(s) for each distribution attached to the dataset	-0.056	0.570	0.611	-0.111	0.901
u68	Validity of format of the email address provided within Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value	-0.083	0.599	0.656	-0.111	0.901

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
i43	Accuracy of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-type header field	-0.111	0.397	0.787	-0.333	0.975
u67	Existence of the Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value	-0.139	0.486	0.792	-0.333	0.975
f33	Existence of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value	-0.139	0.453	0.808	0.111	0.733
c9	Existence of the Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime value	-0.139	0.377	0.849	0.111	0.733
c12	Validity of format of the date provided within Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date value (e.g., according to ISO 8601)	-0.167	0.395	0.879	0.111	0.733
f36	Conformity of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier	-0.167	0.177	0.989	-0.111	0.901
u69	Validity of format of the HTTP URL provided within Stewardship_Record [Data_Set] ->contact; Cataloged_Resource ->contact_point value	-0.194	0.429	0.895	-0.111	0.901
f35	Existence of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset	-0.194	0.391	0.913	-0.333	0.975
u71	Existence of the Data_Set_Provenance ->generation_type; Dataset ->was_generated_by value	-0.222	0.458	0.908	-0.111	0.901
u70	Existence of the Stewardship_Record [Data_Set] ->organization; Organization / Person | foaf:Organization ->foaf:name value	-0.278	0.404	0.963	-0.111	0.901

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f27	Intrinsic precision of the text provided within Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is determined by spelling mistakes	-0.500	0.451	0.995	-0.778	1.000
f26	Readability of the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is computed by using the Flesch-Kincaid Reading Ease test	-0.639	0.220	1.000	-1.000	1.000
f28	Readability of the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is computed by using the Flesch-Kincaid Reading Ease test	-0.667	0.280	1.000	-1.000	1.000

Table D.5: Mean Q sort rank ($\bar{x}$) and Lawshe's content validity ratio (CVR) together with the corresponding p-values based on the t-test and the exact binomial test (BT) for the individual indicators of the contextuality dimension in the initial version of the theoretical framework.

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f22	Existence of the Data_Set ->spatial_coverage; Dataset ->spatial / geographical_coverage value	0.511	0.302	0.000	1.000	0.007
c18	Existence of the Data_Set_Provenance ->originator; Cataloged_Resource ->creator value	0.467	0.300	0.001	0.778	0.055
c2	The timely Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value is determined in relation to the Data_Set ->accrual_periodicity; Dataset ->frequency value	0.444	0.328	0.002	1.000	0.007
c11	Existence of the Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date value	0.444	0.260	0.000	1.000	0.007
c13	Existence of the Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value	0.400	0.447	0.014	0.556	0.194

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f31	Existence of the Designation [Data_Set] ->sign; Cataloged_Resource ->title value	0.356	0.410	0.016	0.556	0.194
c9	Existence of the Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime value	0.311	0.362	0.016	0.778	0.055
c4	Existence of the Data_Set_Distribution ->issued_date; Distribution ->release_date value for each distribution attached to the dataset	0.244	0.260	0.011	0.778	0.055
f21	Existence of the Classification [Data_Set] ->classification_scheme_name; Cataloged_Resource ->theme / category value	0.244	0.606	0.131	0.778	0.055
u56	Existence of the Data_Set ->rights; Cataloged_Resource ->license value	0.178	0.406	0.112	0.556	0.194
c17	Existence of the Registration_State [Data_Set] ->effective_date; Cataloged_Resource ->release_date value	0.156	0.445	0.162	0.333	0.430
c3	Existence of the Data_Set ->rights; Cataloged_Resource ->rights value	0.133	0.490	0.219	0.111	0.688
c12	Validity of format of the date provided within Administered_Item [Data_Set] ->last_change_date; Catalog_Record ->update / modification_date value (e.g., according to ISO 8601)	0.111	0.226	0.089	0.556	0.194
c1	Existence of the Data_Set ->accrual_periodicity; Dataset ->frequency value	0.067	0.346	0.290	0.556	0.194
f34	Conformity of the Scoped_Identifier [Data_Set] ->identifier; Cataloged_Resource ->identifier value with an identifier schema (e.g., URN, DOI, trusty URI) to ensure the uniqueness of an identifier	0.022	0.578	0.456	-0.111	0.875
c19	Existence of the Data_Set_Quality_Assessment ->statement; dqv:QualityAnnotation ->oa:hasBody value	0.000	0.678	0.500	0.111	0.688

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
f30	Semantic distance between the Designation [Data_Set_Distribution] ->sign; Distribution ->title value for each distribution attached to the dataset AND the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset	-0.022	0.636	0.540	0.333	0.430
i45	Openness of the Data_Set_Distribution ->media_type; Distribution ->media_type value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	-0.022	0.291	0.588	0.111	0.688
c15	Existence of the N/A; Distribution ->update / modification_date value for each distribution attached to the dataset	-0.044	0.410	0.623	0.333	0.430
c5	Validity of format of the date provided within Data_Set_Distribution ->issued_date; Distribution ->release_date value for each distribution attached to the dataset (e.g., according to ISO 8601)	-0.067	0.265	0.764	0.333	0.430
c10	Validity of format of the date provided within Data_Set_Provenance ->issued_date; prov:Entity ->prov:generatedAtTime value (e.g., according to ISO 8601)	-0.067	0.224	0.801	0.111	0.688
u62	Conformity of the Data_Set_Distribution ->rights; Distribution ->license value for each distribution attached to the dataset with one of the licenses from the predefined list provided by the Open Definition related to licenses	-0.067	0.400	0.685	0.111	0.688
i41	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed machine-readable file formats	-0.067	0.300	0.738	-0.111	0.875
c14	Validity of format of the date provided within Administered_Item [Data_Set] ->last_change_date; Cataloged_Resource ->update / modification_date value (e.g., according to ISO 8601)	-0.089	0.302	0.799	-0.333	0.966

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
u57	Conformity of the Data_Set ->rights; Cataloged_Resource ->license value with one of the licenses from the predefined list provided by the Open Definition or the EU Vocabularies related to licenses	-0.089	0.285	0.812	0.111	0.688
f35	Existence of the Scoped_Identifier [Data_Set_Distribution] ->identifier; N/A value for each distribution attached to the dataset	-0.111	0.333	0.827	0.111	0.688
i40	Openness of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is checked based on a predefined set of confirmed open / non-proprietary formats	-0.156	0.397	0.863	-0.111	0.875
i46	Accuracy of the Registration_Authority [Data_Set] ->documentation_language_identifier; Cataloged_Resource ->language value is computed by using language detection on the actual resource and / or HTTP content-language header field	-0.156	0.357	0.886	-0.333	0.966
i47	Conformity of the Registration_Authority [Data_Set] ->documentation_language_identifier; Cataloged_Resource ->language value to a given standard such as ISO 639	-0.156	0.433	0.844	-0.111	0.875
c6	Existence of the Data_Set_Distribution ->rights; Distribution ->rights value for each distribution attached to the dataset	-0.178	0.429	0.875	0.111	0.688
u60	Conformity of the Data_Set_Distribution ->rights; Distribution ->access_rights value for each distribution attached to the dataset with the EU controlled vocabulary for access rights	-0.200	0.387	0.920	-0.333	0.966
f28	Readability of the Definition [Data_Set_Distribution] ->text; Distribution ->description value for each distribution attached to the dataset is computed by using the Flesch-Kincaid Reading Ease test	-0.267	0.539	0.912	-0.556	0.994

Id	Individual indicator	$\bar{x}$	SD	p t-test	CVR	p BT
r49	Validity of format of the HTTP URL provided within Data_Set_Distribution ->access_url; Distribution ->access_URL value for each distribution attached to the dataset	-0.289	0.470	0.949	-0.333	0.966
c16	Validity of format of the date provided within N/A; Distribution ->update / modification_date value for each distribution attached to the dataset (e.g., according to ISO 8601)	-0.333	0.300	0.995	-0.333	0.966
i38	Accuracy of the Data_Set_Distribution ->format; Distribution ->format value for each distribution attached to the dataset is computed by using file-extension of the actual resource and / or by taking the format specified in the HTTP content-type header field	-0.333	0.224	0.999	-0.778	1.000
c7	Existence of the Data_Set_Distribution ->size; Distribution ->byteSize value for each distribution attached to the dataset	-0.378	0.543	0.965	-0.556	0.994
c8	Accuracy of the Data_Set_Distribution ->size; Distribution ->byteSize value for each distribution attached to the dataset is computed by using the information specified in the HTTP content-length header field	-0.467	0.490	0.989	-0.333	0.966
f26	Readability of the Data_Set | Definition [Data_Set] ->comments | text; Cataloged_Resource ->description value is computed by using the Flesch-Kincaid Reading Ease test	-0.533	0.600	0.986	-0.778	1.000

CURRICULUM VITAE

Barbara Šlibar, born on 14 October 1991 in Varaždin, is a graduate of the Faculty of Organization and Informatics at the University of Zagreb (UniZG FOI). She received the Rector's Award for her research work in the 2015/2016 academic year and the Rotary Club Varaždin Excellence Award for her master's thesis in 2016.

After working as an HR manager in an IT company, Barbara joined UniZG FOI in 2018. In the same year, she enrolled on the postgraduate doctoral study programme in Information Science. Following a decision by the Faculty Council on 2 April 2020, Barbara was appointed an associate and teaching assistant for courses related to business decision-making.

In addition to her involvement in teaching and related activities, Barbara actively participates in research projects – Customer Experience of the Future – Smart Specializations and Modern Communication and Collaboration Technologies (completed), Digital and entrepreneurial skills for European teachers in the COVID-19 world (completed), Improving HEI maturity to implement learning analytics (in progress).

Barbara currently serves as a representative of employees in associate positions on both the Faculty Council and the Council of the Social Sciences and Humanities Area. She has been a member of the UniZG FOI Laboratory for Strategic Planning and Decision Making, the Croatian Society for Operational Research and the Croatian Society for Information, Communication and Electronic Technology. Barbara was also a member of the UniZG FOI Student Council for two terms.

Her participation in EU DataViz 2021, where she was the only Croatian representative, was a remarkable success. This conference, which focused on open data and its visualisation, kicked off the very first EU Open Data Days. Together with Professor Enrique Mu (Carlow University, USA), she presented the results of their joint research under the title 'The Current State of National Metadata - How to Unlock Open Data Benefits'. She also received the 2022 Annual Award to Young Scientists and Artists from the Society of University Teachers and other Scientists in Zagreb.

Published research

1. Gusić Munđar, J., Rako, S., Šimić, D., & Šlibar, B. (2021). IDENTIFYING COMMUNITIES OF PRACTICE IN LEARNING ANALYTICS AND EDUCATIONAL DATA MINING USING TOPIC MODELLING AND SOCIAL NETWORK ANALYSIS. In *ICERI2021 Proceedings* (pp. 6033–6039). Online Conference: International Academy of Technology, Education and Development (IATED). doi: 10.21125/iceri.2021.1363

2. Gusić Munđar, J., Rako, S., & Šlibar, B. (2022). Finding the most representative Latent

Dirichlet Allocation run for topic modelling. In (pp. 117–117). Šibenik, Croatia: Croatian Operational Research Society (CRORS).

3. Kadoić, N., & Šlibar, B. (2020). Increasing the Learning Efficiency in Decision-Making Field using the Workshop Activity in Moodle. In *Proceedings of the 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics* MIPRO 2020 (pp. 592–597). Opatija, Croatia: Hrvatska udruga za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku - MIPRO.

4. Kutnjak, A., Hrustek, L., Šlibar, B., Pihir, I., Furjan, M. T., Hrustek, N. Ž., . . . Mekovec, R. (2021). Insight of Croatian Open Data Portals Functionalities According to TODO Interdisciplinary Assessment Framework v2.0. In *Proceedings of Central European Conference on Information and Intelligent Systems* (pp. 351–357). Varaždin, Croatia: Faculty of Organization and Informatics, University of Zagreb.

5. Lovrenčić, S., Plantak Vukovac, D., Šlibar, B., Nahod, B., Andročec, D., Šestak, M., & Stapić, Z. (2018). Igrifikacija: prema sistematizaciji termina na hrvatskom jeziku. In *ZBORNIK RADOVA RAČUNALNE IGRE 2018 - STRUČNA KONFERENCIJA* (pp. 1–12). Varaždin, Croatia: Faculty of Organization and Informatics, University of Zagreb.

6. Novak, M., Šlibar, B., Kermek, D., & Begičević, N. (2020). EVALUATION OF FACTOR ANALYSIS METHODS: STUDENT FEEDBACK CASE STUDY. In *INTED2020 Proceedings* (pp. 8208–8216). Valencia, Spain: International Academy of Technology, Education and Development (IATED). doi: 10.21125/inted.2020.2241

7. Rako, S., Šimić, D., Šlibar, B., & Gusić Munđar, J. (2022). Improving stability of detected topics in learning analytics. In *Proceedings of the 45th Jubilee International Convention on Information and Communication Technology, Electronics and Microelectronics* MIPRO 2021 (pp. 584–588). Opatija, Croatia: Hrvatska udruga za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku - MIPRO. doi: 10.23919/MIPRO55190.2022.9803778

8. Zlatić, L., Šlibar, B., & Begičević Ređep, N. (2021, September). Decision Making Styles in Higher Education Institutions: Systematic Literature Review. In *Proceedings of the 44th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2021* (pp. 826–832). Opatija, Croatia: Hrvatska udrugaza informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku - MIPRO. doi: 10.23919/MIPRO52101.2021.9596666

9. Šlibar, B. (2019a). Framework for Quality Assessment of Open Datasets. In *Proceedings of the 11th International Doctoral Seminar (IDS 2019)* (pp. 19–21). Varaždin, Croatia: Faculty of Organization and Informatics and Faculty of Materials Science and Technology in Trnava.

10. Šlibar, B. (2019b, November). Predicting the Number of Downloads of Open Datasets by Naive Bayes Classifier. *TEM JOURNAL-TECHNOLOGY EDUCATION MANAGEMENT INFORMATICS, 8*(4), 1331–1338. doi: 10.18421/TEM84-33

11. Šlibar, B. (2020). Modeling Open Data Usage: Decision Tree Approach. In *Proceedings of ICICC 2019* (Vol. 1087, pp. 57–64). Ostrava, Czech Republic: Springer. doi: 10.1007/978-981-15-1286-5_6

12. Šlibar, B., Gusić Munđar, J., Rako, S., & Šimić, D. (2022, March). Co-occurrence patterns of issues and guidelines related to ethics and privacy of learning analytics in higher education—literature review. In *LAK22 Conference Proceedings* (pp. 577–582). Virtual, Online: Association for Computing Machinery. doi: 10.1145/3506860.3506974

13. Šlibar, B., & Mu, E. (2021, November). The current state of national metadata - how to unlock open data benefits. Online Conference. Retrieved from *https://op.europa.eu/documents/7950956/9399502/EU-DataViz-2021_23-November_Barbara-Slibar.pdf/d60b0b9f-b6b3-e645-3551-5434fce9e6ad?t=1638458673734*

14. Šlibar, B., & Mu, E. (2022, October). OGD metadata country portal publishing guidelines compliance: A multi-case study search for completeness and consistency. *Government Information Quarterly, 39*(4). doi: 10.1016/j.giq.2022.101756

15. Šlibar, B., Oreški, D., & Begičević Ređep, N. (2021, April). Importance of the Open Data Assessment: An Insight Into the (Meta) Data Quality Dimensions. *SAGE Open, 11*(2), 1–18. doi: 10.1177/21582440211023178

16. Šlibar, B., Oreški, D., & Klačmer Čalopa, M. (2023, May). Push and pull factors in brain drain among university students. Management: *Journal of Contemporary Management Issues, 28*(1), 65–80. doi: 10.30924/mjcmi.28.1.5

17. Šlibar, B., Oreški, D., & Kliček, B. (2018). Aspects of Open Data and Illustrative Quality Metrics: Literature Review. In (pp. 90–99). Lisbon, Portugal: Varazdin Development and Entrepreneurship Agency.

18. Šlibar, B., Plantak Vukovac, D., Lovrenčić, S., Šestak, M., & Andročec, D. (2018). Gamificationin a Business Context: Theoretical Background. In *Proceedings of Central European Conference on Information and Intelligent Systems* (pp. 123–131). Varaždin, Croatia:Faculty of Organization and Informatics, University of Zagreb.

19. Šlibar, B., Zlatić, L., & Begičević Ređep, N. (2021). ETHICAL AND PRIVACY ISSUES OF LEARNING ANALYTICS IN HIGHER EDUCATION. In *ICERI2021 Proceedings* (pp. 3064–3074). Online Conference: International Academy of Technology, Education and Development (IATED). doi: 10.21125/iceri.2021.0761