
 <p>Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration</p>
[JANUARY 2023]	Advancing Open Scholarship
	D8.5 – GUIDELINES ON THE RESEARCH DATA IN THE HUMANITIES Version 1.0 – Final/PUBLIC
	H2020-INFRAEOSC-2019 Grant Agreement 863420

The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863420

Disclaimer - “The content of this publication is the sole responsibility of the TRIPLE consortium and can in no way be taken to reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains.”

This deliverable is licensed under a Creative Commons Attribution 4.0 International License



Project Acronym:	TRIPLE
Project Name:	Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration
Grant Agreement No:	863420
Start Date:	1/10/2019
End Date:	31/03/2023
Contributing WP	WP8, WP2
WP Leader:	Max Weber Stiftung
Deliverable identifier	D8.5
Contractual Delivery Date: 31/12/2022	Actual Delivery Date: 20/01/2023
Version:	1.0 Final
Dissemination level	PU

Revision History

Version	Created/Modifier	Comments
0.1	Tomasz Umerle (IBL PAN) – editor; Marta Błaszczńska (IBL PAN), Magdalena Wnuk (IBL PAN), Mateusz Franczak (IBL PAN), Jadranka Stojanovski (University of Zadar), Cezary Rosiński (IBL PAN), Nikodem Wołczuk (IBL PAN), Agnieszka Mikołajczyk-Bareła (Voicelab), Agnieszka Karlińska (IBL PAN), Maciej Ogrodniczuk (IPI PAN), Piotr Pęczik (UŁ), Bianca Kramer (Sesame Open Science, Open Abstracts), Silvio Peroni (University of Bologna, OpenCitations), Luca De Santis (Net7)	First full version
0.2	Luca De Santis (Net7), Lorna Balkan (CESSDA), Ana Inkret, Agnieszka Karlińska, Nikodem Wołczuk, Cezary Rosiński, Tomasz Umerle, Jadranka Stojanovski, Karolina Przysiecka, Maciej Maryl (IBL PAN), Erzsébet Tóth-Czifra (DARIAH)	Modified full version
1	Tomasz Umerle (IBL PAN) – editor; Marta Błaszczńska (IBL PAN), Magdalena Wnuk (IBL PAN), Mateusz Franczak (IBL PAN), Jadranka Stojanovski (University of Zadar), Cezary Rosiński (IBL PAN), Nikodem Wołczuk (IBL PAN), Agnieszka Mikołajczyk-Bareła (Voicelab), Agnieszka Karlińska (IBL PAN), Maciej Ogrodniczuk (IPI PAN), Piotr Pęczik (UŁ), Bianca Kramer (Sesame Open Science, Open Abstracts), Silvio Peroni (University of Bologna, OpenCitations), Luca De Santis (Net7).	First released version

Table of contents

Executive summary	7
1. Aim of the deliverable	8
2. Research data in the humanities	9
2.1 Specificity of the humanities	9
2.2 Humanities metadata as research data	11
3. Persistent identifiers	12
3.1. PIDs the SSH – Current State and Upcoming Challenges	12
3.2. GoTriple, Dublin Core and Persistent Identifiers	20
3.3. Guidelines for PIDs as humanities research data in the context of TRIPLE – summary	25
4. Keywords	26
4.1. Keywords in the SSH – current state and upcoming challenges	28
4.2. Keyword generation in SSH	36
4.3. Keywords in the TRIPLE Project	38
4.4. Guidelines for keywords as humanities research data	41
5. Abstracts	42
5.1. Abstracts in the SSH	42
5.2. Highway to abstract. The present and the future of automatically generated abstracts	54
5.3. Guidelines for abstracts as humanities research data	61
6. Citations	62
6.1. Transparency meets open citations	62
6.2. Citation data and GoTriple’s data providers	67
6.3. Guidelines for citations as humanities research data	70
7. Conclusion	71
Bibliography	72
Appendix: Data Enrichment in TRIPLE: the current state of work	76

List of Figures

- Figure 3.1.1. Entities in the publishing system (p. 14).
- Figure 3.2.1. PIDs from Library of Science in JATS format (bibliotekanauki.pl, September 2022) (p. 21).
- Figure 3.2.2. Approaches to store PIDs in DC considered by DCMI (p. 22).
- Figure 3.2.4. Using tag attributes to store PIDs (p. 22).
- Figure 4.1.1. Keywords typology (p. 27).
- Figure 4.2.1. Use of automatic keyword extraction tools by editors of Polish scientific journals and publishers (p. 35).
- Figure 4.2.2. Interest among editors of Polish scientific journals in using new automatic keyword extraction software (p. 36).
- Figure 4.2.3. Interest among Polish scientific publishers in using new automatic keyword extraction software (p. 36).
- Figure 4.3.2. The GoTriple Vocabulary in Semantics.gr (p. 39).
- Figure 5.1.1. Abstract coverage in Crossref per publication type (publication years 2020-2022) (p. 45).
- Figure 5.1.2. Abstract coverage in Crossref for journal articles – selected publishers (publication years 2020-2022) (p. 46).
- Figure 5.1.3. Abstract coverage in Crossref for books, book chapters and monographs – selected publisher (publication years 2020-2022) (p. 47).
- Figure 5.1.4. Abstract coverage in Crossref for books, book chapters and monographs separately - Oxford University Press (OUP) and Cambridge University Press (CUP) (publication years 2020-2022) (p. 48).
- Figure 5.1.5. Abstract coverage in Crossref for books, book chapters and monographs separately - University of Warsaw and Uniwersytet Lodzki (publication years 2020-2022) (p. 48).
- Figure 5.1.6. Abstract coverage in Crossref for books, book chapters and monographs separately - Berghahn Books, Amsterdam University Press (AUP) and Bologna University Press (publication years 2020-2022) (p. 49).
- Figure 5.1.7A, 5.1.7B. Abstract coverage for Crossref DOIs in Crossref and OpenAlex (A) and for publications with or without DOI in OpenAlex (B) per publication type (publication years 2020-2022, sampled June 2022) (p. 52).
- Figure 5.2.1. Articles with abstracts in Library of Science (bibliotekanauki.pl, September 2022) (p. 54).
- Figure 5.2.2. Records with abstracts in Digital Repository of Scientific Institutes RCIN (rcin.org.pl, September 2022).
- Figure 5.2.3. Supplementing missing abstract workflow (p. 55).
- Figure 5.2.4. Interest among editors of Polish scientific journals in using new automatic metadata extraction software (p. 59).
- Figure 5.2.5. Interest among Polish scientific publishers in using new automatic metadata extraction software (p. 59).
- Figure 5.2.6. Interest among editors of Polish scientific journals in using new abstract generation tools (p. 60).
- Figure 5.2.7. Interest among Polish scientific publishers in using new abstract generation tools (p. 60).
- Figure 6.1.1. Some initiatives pushing for reforming the principles behind research assessment systems (p. 64).
- Figure 6.1.2. Collaborations between OpenCitations and other Open Science infrastructures and services (p. 65).
- Figure 6.2.1. Citations in Library of Science in JATS format (bibliotekanauki.pl, September 2022) (p. 67).

Figure 6.2.2. Missing citations - possible solutions (p. 68).

Figure 6.2.3. Example of automatic parsing using Annif software (p. 69).

Figure Appendix 1. Publications data flow (p. 77).

Figure Appendix 2. PID URL redirection to a SEO and user friendly URL (p. 80).

List of tables

Table 4.2.1. Overview of the results of the qualitative evaluation of three approaches to keyword extraction for Polish (p. 33).

Table 4.3.1. Progress in enhancing multilingualism in GoTriple Vocabulary (p. 39).

List of acronyms (not explicitly decoded in the main text)

APC	Article Publishing Charge
API	Application Programming Interface
BASE	Bielefeld Academic Search Engine
BERT	Bidirectional Encoder Representations from Transformers
BITS	Book Interchange Tag Suite
DL	Deep Learning
FAIR	Findable, Accessible, Interoperable, Reusable
GLAM	Galleries, Libraries, Archives, Museums
GPT	Generative pre-training
ISO	International Organization for Standardization
JATS	Journal Article Tag Suite
JSON	JavaScript Object Notation
ML	Machine Learning
MORESS	Mapping of Research in European Social Sciences and. Humanities
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCR	Optical Character Recognition
OLR	Optical Layout Recognition
OPERAS	European Research Infrastructure for the development of open scholarly communication in the social sciences and humanities
RISIS	Research Infrastructure for Science, technology and Innovation policy Studies
SEO	Search Engine Optimization
SSH	Social Sciences and Humanities
STEM	Science, Technology, Engineering, Mathematics
TRIPLE	Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration
UNESCO	United Nations Educational, Scientific and Cultural Organization
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UTF	Unicode Transformation Format
XML	Extensible Markup Language

Executive summary

This report focuses on the metadata as a specific type of research data in the humanities by analysing key metadata elements – persistent identifiers (PIDs), abstracts, keywords and citations. It defines those elements, outlines challenges for processing them in the humanities and presents the challenges for GoTriple as the metadata aggregator of this kind of research data.

The assumption is that GoTriple is a specific kind of research dataset on its own that can and will be reused by stakeholders such as other metadata aggregators, indexers, publishers, information services (i.e. providers of scholarly metrics), but also scientists interested in data-driven research (cultural analytics, scientometrics, bibliometrics, etc.). This demands a good understanding of key metadata elements important to GoTriple's aggregation and enrichment processes (abstracts, keywords) and their development (PIDs, citations).

Chapter 1 defines the aim of the deliverable, context of its creation and its audience.

Chapter 2 discusses the specificity of the research data in the humanities and this report's position in the rich discussions on the topic.

Chapter 3 – dedicated to PIDs – presents the overview of the topic and the challenges related to the PID's uptake by the humanities, such as the role of cultural heritage data for the humanities, importance of bibliodiversity and multilingualism (subchapter 3.1), then it proceeds to the discussion of processing PIDs from GoTriple's data providers by focusing on data dispersion and heterogeneity (subchapter 3.2).

Chapter 4 – dedicated to keywords – begins with the typology of keywords and the expected standards they should adhere to (subchapter 4.1). Subchapter 4.2 tackles the issue of automated generation of keywords and proposes different approaches applicable in the context of GoTriple. In the subchapter 4.3 a current approach to keyword organisation in GoTriple is presented, with focus on the GoTriple vocabulary that responds to the need for keywords LOD-ification and can be in the future reused for automated keyword generation.

Chapter 5 – dedicated to abstracts – starts with the comprehensive presentation of the abstract ecosystem, offering also a specific perspective on SSH. Subchapter 5.2 offers solutions to the issues of “missing abstracts” which are aimed at the needs of the GoTriple platform.

Chapter 6 – dedicated to citations – offers an overview of the topic and its relevance to the SSH. In the subchapter 6.2 an analysis of issues related to GoTriple's expression of citation data is presented (that relates especially to the challenge of processing different citation formats and citation data quality).

Each chapter concludes with a summary of the guidelines for the specific metadata type for the humanities.

1. Aim of the deliverable

This report, *Guidelines on the research data in the humanities*, gives a perspective on the research data by focusing on **metadata¹ as research data in the humanities** and treating **GoTriple – a metadata aggregator – as a unique humanities research data collection on its own** which is being constructed through aggregation, normalisation and enrichment procedures and made available for machine exploitation via APIs².

The GoTriple platform collects, enriches and provides access to data which could and should be further re-used for building new data services providing insights into the humanities' scientific output³. These insights from GoTriple's dataset analysis might be produced through scientometrics, bibliometrics, cultural analytics and other quantitative and data-driven research methods and/or used by diverse stakeholders in tools and services providing statistics, metrics or visualisations (which can be of interest to funding bodies, libraries, other GLAM institutions and policy makers). To make this kind of reuse possible, it is important to recognise the GoTriple dataset – and the data it processes – as research data in their own right and investigate their broader contexts, both intellectual and technological.

To further develop this dimension of the GoTriple platform it is important to investigate critical metadata elements of the GoTriple dataset and its development in two dimensions: their role in the scholarly ecosystem and their specificity in the context of the TRIPLE project.

This deliverable focuses on four important types of metadata – **persistent identifiers [PIDs], abstracts, keywords and citations** – and presents their specificity in the context of the humanities (and, more broadly, SSH), discussed in this report by both TRIPLE project contributors and external experts.

Based on these comprehensive discussions this deliverable offers a set of guidelines for defined metadata types, focusing on humanities and GoTriple's challenges in their processing and reuse. Each chapter – devoted to one metadata type – is concluded with a summary of those guidelines which outline the main tasks and challenges that need to be addressed to leverage the potential of a certain data type. Thanks to that, a set of guidelines are delivered that could guide any party interested in building a rich metadata-based research dataset and which will contribute to further discussions on the development of the GoTriple dataset. This deliverable shall be of particular interest to actors such as curators of large metadata repositories, aggregators, publishers, indexes and registries and information services (metrics providers etc.).

This deliverable has been prepared partly through the booksprint *The role of open metadata in the SSH scholarly communication* organised by IBL PAN on September 7-9, 2022⁴. During this event external experts collaborated with TRIPLE's contributors to provide inputs into this deliverable by exchanging knowledge on the four key metadata types (PIDs, keywords, abstracts and citations) as examples of research data in the humanities.

¹ Metadata are structured information about the form, content, and context of documents in any form (textual, graphic, musical notation, etc.) or medium (printed, electronic, etc.).

² L. De Santis. TRIPLE Deliverable: D6.6 API's Development -RP3 (Draft). Zenodo. 2022. <https://doi.org/10.5281/zenodo.7371832>.

³ See Appendix 1 below for an overview of the current data enrichment workflow in GoTriple.

⁴ See <https://project.gotriple.eu/events/triple-booksprint/>

2. Research data in the humanities

Marta Błaszczńska, Mateusz Franczak, Magdalena Wnuk

Data in social sciences and humanities (SSH) has been recognised and growing in importance over the last few years and the TRIPLE project has been no exception, with data forming a significant part of the discussions, penetrating decisions about the project's foundations, planned actions and sustainability. Before delving into the details of data and metadata within the GoTriple platform, it is important to take a brief look at the history of data in the context of SSH more generally and humanities more specifically to explore (some of) the main themes, challenges and opportunities that the discourses have brought.

2.1 Specificity of the humanities

Since the nature of data and thus attitudes towards them differ between social sciences and humanities, it is often within the latter that most heated debates have been taking place. It is rare that one needs to persuade a quantitative sociologist that the numbers they have been collecting for her last paper are in fact data. On the other hand, humanists have been weary of using the word 'data' to describe the resources they have been collecting, producing, analysing and publishing within their academic workflows. Indeed, it has often been the language that proved to be the biggest barrier in bringing data into the humanities, where the specifically selected semantic choices such as 'primary sources, secondary sources, theoretical documents, bibliographies, critical editions, annotations, notes'⁵ have been discouraging scholars from describing their research processes as simply 'data'.

At the same time and partly stemming from this, there is often a fear of simplification or missing out on important contexts and nuances that enter the humanities discussions together with the concept of data. For example, big data approaches are frequently accused of being reductionist and underrepresenting the richness of information⁶. Moreover, historians or literary scholars often would perceive their work as interpretive rather than data-driven. To speak in terms of data may seem not to capture the whole nature of their work. Indeed, it may also be the nature of humanities, where the relationship between the scholar and the resources that they may be studying is different to natural sciences. Jennifer Edmond and Erzsébet Tóth-Czifra stress that, therefore, the issue is also a 'material one'. Persons conducting humanities research rarely produce data, they often do not really own it, basing their work on historical and cultural foundations of existing interpretations⁷.

So what should be understood as data in the humanities? Definitions and typologies differ and indeed are (and should be) discipline- and workflow specific⁸. Importantly,

⁵ J. Edmond, E. Tóth-Czifra, "Open Data for Humanists, A Pragmatic Guide", 10 December 2018: 1, <https://doi.org/10.5281/zenodo.2657248>.

⁶ J. Edmond, N. Horsley, R. Kalnins, J. Lehman, M. Priddy, T. Stodulka, "Big Data & Complex Knowledge. Observations and Recommendations for Research from the Knowledge Complexity Project", K-PLEX. University College Dublin, 2018: 8-9. https://kplexproject.files.wordpress.com/2018/04/trinity-big-data-report-jklr_04.pdf.

⁷ J. Edmond, E. Tóth-Czifra. 'Open Data...', p. 1.

⁸ See B. Gualandi, L. Pareschi, S. Peroni. 'What Do We Mean by "Data"? A Proposed Classification of Data Types in the Arts and Humanities'. arXiv, 15 July 2022. <https://doi.org/10.48550/arXiv.2205.06764> for a proposal of 13 data types based on the interviews with researchers in philology and literary criticism, language and linguistics, history of art, computer

providing a comprehensive answer to this complex question is not the aim of this deliverable. Several general specificities of humanities data ought to be highlighted here (in the light of metadata considerations that come later in this document), however. Firstly, an aspect especially relevant to the TRIPLE project, data in humanities – in addition to scholarly sources – often include cultural heritage resources. Thus, collaboration with the GLAM (galleries, libraries, archives, museums) sector needs to be established and fostered for successful studies to be carried out⁹.

Secondly, the importance of multilingualism in humanities cannot be underestimated. As it has often been argued, topics and issues that are important to national and local cultural phenomena often ought to be discussed in the languages that are understandable to the persons who are most directly affected. The same goes for data on a more general level – they may sometimes be translated into English or other languages but a strong understanding of the local context is needed for analysis and interpretation¹⁰.

Thirdly, data standards and general guidelines need to always be reevaluated in the humanities context. For instance, while the FAIR principles (for data to be findable, accessible, interoperable and reusable) have entered the humanities a while back, discussions on what being FAIR means need to be constantly reassessed in terms of humanities resources¹¹ and will, for example, be understood differently in the context of critical literary editions as compared to philosophical sources or anthropological notes. On the other hand, straightforward reproducibility of studies is often seen as irrelevant in humanities because – as we mentioned above – a large amount of research relies on interpretative work¹². However, it is still important to allow our readers to have as full an

science, and archival studies, conducted at the Department of Classical Philology and Italian Studies at the University of Bologna. See also M. Maryl, M. Błaszczewska, B. Szleszyński, T. Umerle, “Dane badawcze w literaturoznawstwie”, *Teksty Drugie. Teoria literatury, krytyka, interpretacja*, 2 (1 March 2021): 13–44 for an attempt to draw a coherent typology for data in Polish literary studies. The data stories and other discipline-focused work undergoing in the context of DARIAH-EU [Research Data Management Working Group](#) are also worth mentioning in this context (see e.g. E. Tóth-Czifra, N. Truan, “Creating and Analyzing Multilingual Parliamentary Corpora”, 2021. <https://halshs.archives-ouvertes.fr/halshs-03366486>). Its members are currently preparing a publication in the context of the ‘Research Data Management for Arts and Humanities: Integrating Voices of the Community’ project under the DARIAH Working Groups Funding Scheme 2021-2023 that ‘covers and provides practical know-how for both researchers and the new research support professionals (data stewards, subject librarians, open science officers, etc.) working with them.’ (see: <https://www.dariah.eu/2022/05/23/dariah-working-groups-funding-scheme-2021-2023-meet-the-winnig-projects/>).

⁹ G. Angelaki, K. Badzmirowska, D. Brown, V. Chiquet, J. Colla, J. Finlay-McAlester, K. Grabowska, et al., *How to Facilitate Cooperation between Humanities Researchers and Cultural Heritage Institutions. Guidelines*, Warsaw, Poland: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 10 March 2019. <https://doi.org/10.5281/zenodo.2587481>.

¹⁰ For a more detailed discussion on multilingualism in humanities see: A. Balula, L. Caliman, S. Fiorini, S. Jarmelo, D. Leão, P. Mounier, J.-F. Nomine, et al. *Innovative Models of Bibliodiversity in Scholarly Publications: OPERAS Special Interest Group Multilingualism White Paper*, 8 November 2021. <https://doi.org/10.5281/zenodo.5653084>.

¹¹ See: N. Harrower, M. Maryl, T. Biro, Immenhauser, ALLEA Working Group E-Humanities, *Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities*, Berlin: ALLEA - All European Academies, February 2020. Digital Repository of Ireland. <https://repository.dri.ie/catalog/tq582c863>.

¹² There has been a lot of debate on the topic, however. See R. Peels, “Replicability and replication in the humanities”, *Res Integr Peer Rev* 4, 2 (2019). <https://doi.org/10.1186/s41073-018-0060-4>; J. Britt Holbrook, B. Penders, S. de Rijcke, “The humanities do not need a replication drive”, *CWTS Blog* (archive), 21 January 2019. <https://www.cwts.nl/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive>; R.

understanding of the resources we have been using as possible, even though they may have completely different perceptions of their meaning and relevance to our arguments.

2.2 Humanities metadata as research data

While the humanities community has different approaches to what research data mean for their disciplines, it needs to be emphasised that the metadata types that are a point of emphasis in this report – PIDs, keywords, abstracts and citations – are just as valuable types of research data for humanists, as they are for other disciplines (social sciences, but also STEM).

Hence, it is even more important to address critical challenges for humanities research data in these fields, especially when humanities are lagging behind in adhering to state-of-the-art or pose very specific problems. Of course, some of the challenges we will be discussing below are common for the humanities and social sciences (rather than strictly specific to the humanities) and then we discuss them jointly. However, in all cases when the humanities perspective differs and demands a very different approach, we will showcase and detail it, so that the humanities community can benefit from this deliverable.

In the following chapters – each devoted to another metadata type – a comprehensive overview of a topic is presented, with a specific focus on the SSH, and the humanities specifically. Additionally, each chapter discusses these issues on concrete examples from the GoTriple dataset and/or GoTriple’s data providers to facilitate future development of the platform.

Peels, L. Bouter, R. van Woudenberg, “Do the humanities need a replication drive? A debate rages on”, Retraction Watch, 13 February 2019. <https://retractionwatch.com/2019/02/13/do-the-humanities-need-a-replication-drive-a-debate-rages-on/>; J. O’Sullivan, “The humanities have a ‘reproducibility’ problem”, Talking Humanities, 9 July 2019. <https://talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/>, among others.

3. Persistent identifiers

3.1. PIDs the SSH – Current State and Upcoming Challenges

Jadranka Stojanovski¹³

“An identifier is an opaque or explicit number or alphanumeric label which is machine or human readable. It uniquely and permanently identifies and retrieves an object, a document, person, place, organisation, or any entity, in the real world and on the Internet.”

<https://www.ouvri.lascience.fr/open-identifiers-for-open-science/>

Why do we need persistent identifiers (PIDs)?

Persistent Identifiers (PIDs) are unique entity names with the organisational commitment and technical infrastructure to support them indefinitely¹⁴. PID could also be defined as a unique identification code attached to a digital object and registered at an agreed location¹⁵. For PIDs, the resources must be registered in trusted repositories with content that is never changing and which can be referenced and cited this way. Furthermore, these references must be stable, whereas the underlying repositories continuously change hardware, software, physical place or format. It is guaranteed for PID to remain functional, providing access to the resource as it moves from one location to another. PID is not only persistent but actionable: it can be plugged into a web browser and taken to the identified source.

Persistent identifiers (PIDs) provide the long-term preservation of scientific resources to ensure their long-lasting accessibility. PIDs serve as pointers used to identify and retrieve different resources (publications/documents, software, datasets, bibliographic records/metadata files, multimedia materials, and projects/grants) and can also be applied to physical research components, such as research institutions, funders, persons/researchers, samples, artefacts, reagents and instruments. Due to the identifier resolver network, they provide a resolution mechanism enabling direct access to the identified data. They are globally unique, with infinite lifespans, and can be resolved to the physical resource.

According to the UNESCO Science Report (2021), science spending increased worldwide by 19% over the four years, while the number of scientists grew by 13.7% to 8.8 million. In addition, publication output reached 2.9 million articles in 2020, with over 90% of the total from countries with high-income and upper-middle-income economies. Furthermore, the publication output's compound annual growth rate has increased by 5% from 2017 to 2020¹⁶. These numbers do not include other types of research output like preprints, software, research data, and multimedia content, so we can conclude that activities in the research area and their outcomes are growing unstoppably.

Depending on the discipline, scholarly research consists of complex processes such as research planning and design, data creation and collection, data analysis, reporting of findings, dissemination and sharing, and access and reuse. Although the research process steps have not changed, traditional research did not involve a large amount of collected data

¹³ University of Zadar, <https://orcid.org/0000-0001-7399-522X>.

¹⁴ <https://socialhistoryportal.org>

¹⁵ www.ncdd.nl

¹⁶ <https://nces.nsf.gov/pubs/nsb20214>

and required only limited use of technologies. On the other hand, today's research is data-intensive, characterised by a large amount of data, the collection and analysis of which requires complex materials, equipment, infrastructure, tools and software, and new data use and reuse paradigms. In addition, research is interdisciplinary, and research teams are larger, sometimes consisting of big collaboration teams of researchers from different disciplines and countries. In the open science world, the data about the research funder, institution/lab (where the research is carried out), and instruments/equipment is available. Research is preregistered, specifying the research plan before observing the research outcomes. Lab notes, protocols, datasets and publications are shared publicly.

Despite significant development initiated by findability, accessibility, interoperability, and reusability (FAIR) principles, many relations with the research process and its outcomes are lost in the publication-centric landscape of scholarly publishing. Even in cases where data about research infrastructure is available, these data do not have an adequate level of interoperability. Traditional identifiers like International Standard Book Number (ISBN) and International Standard Serial Number (ISSN), providing unique and persistent identifiers for a specific type of resources, were designed for printed resources and are not actionable on the Internet nor interpreted as hyperlinks by web browsers. Therefore, new identifiers were developed for digital data items to provide persistent links to the resources. In practice, the persistent identifier is mapped to up-to-date locators, facilitating access to the physical manifestation of the resource¹⁷. Based on principles of the level of indirection (separating the name from the particular instance addressed) offered by resolution¹⁸ to the "landing page" and/or resource itself, and the use of metadata at the registry level to describe the objects identified, appropriate holders are provided for all services securing reliable interoperability. Appropriate use of PIDs supports the discovery of digital resources, citations, reuse, interoperability and collaboration across facilities, disciplines, institutions and countries, evaluation of impact through citation tracking, trust, efficiency, scalability and innovation¹⁹.

A successful and trustworthy PID system should be built on four pillars:

- Identifier independence of any particular technology or organisation.
- Delivering essential PID functions: a) issuing identifiers (uniqueness, ownership, editable metadata), b) storing identifiers (scalability, integrity, interpretability, versioning), and c) resolving identifiers.
- Separation from data delivery: direct access vs landing web page
- Employing policies for change, including technology change, social change, identifier abandonment, financial sustainability, and decommissioning²⁰.

By using persistent identifiers, relations of publication with authors, affiliations, and parent publications (e.g. journals) are mainly established. However, the relations with software, research data, protocols, references, other versions of the paper (e.g. preprints), type of paper, Article Processing Charges (APC) data, the project within which the research was

¹⁷ J. Hakala, Persistent identifiers: an overview. *KIM Technology Watch Report*. 2010, <http://www.persid.org/downloads/PI-intro-2010-09-22.pdf>.

¹⁸ N. Paskin, "Digital object identifier (DOI®) system", *Encyclopedia of library and information sciences*, 3 (2010): 1586-1592. http://0-www.doi.org.oasis.unisa.ac.za/topics/020210_CSTI.pdf.

¹⁹ The PID Forum. Why Use Persistent Identifiers?, 2019, <https://pidforum.org/t/why-use-persistent-identifiers/714>.

²⁰ N. Car, P. Golodoniuc, J. Klump, "The challenge of ensuring persistency of identifier systems in the world of ever-changing technology", *Data Science Journal*, 16 (2017): 13. DOI: <http://doi.org/10.5334/dsj-2017-013>.

conducted, funder and most importantly, data on the peer review process, reviewers and their reports are mostly lost (Figure 3.1.1). Therefore, they require a broader application of persistent identifiers, supported by the first FAIR principle: “(Meta)data are assigned a globally unique and persistent identifier”.

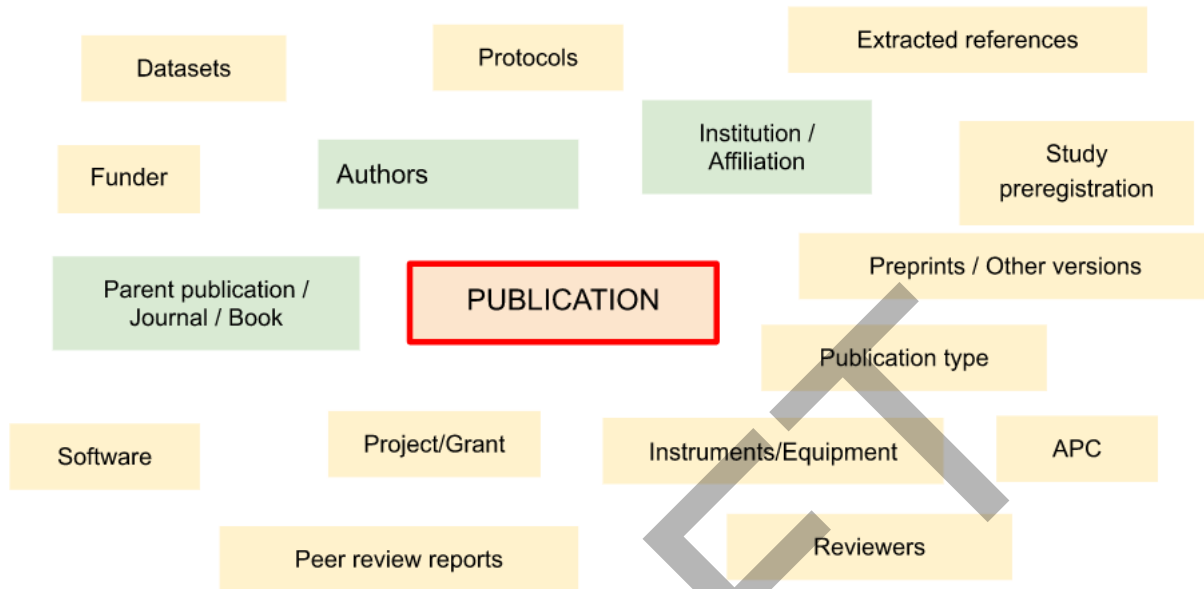


Figure 3.1.1. Entities in the publishing system

Persistent identifiers were first mentioned in the Data Seal of Approval (DSA) issued by Data Archiving and Networked Services (DANS) in 2008, representing a certification system with 16 guidelines for increasing trust in repositories. The discussion has been continued, and a set of criteria further developed. Finally, in 2016 FAIR principles were published, summarising the importance of persistent identifiers²¹. According to these principles, data should be “Findable” online using a persistent identifier (PID) that enables citation and data tracking. Metadata, or information about the data, must be “Accessible.” To be “Interoperable” with other data, the data must be in widely accepted file formats, preferably open file formats, and be characterised using standard vocabularies. The data can be made “Reusable” for other researchers by being accompanied by the appropriate documentation and a user licence, promoting collaboration and maximising the effect of the study outcomes²².

3.1.1 Overview of widely adopted PIDs for published content, authors and institutions

Machine-readable PIDs such as DOI, ORCID and ROR are widely accepted, representing valuable advantages in enabling information sharing across systems.

²¹ P. Wittenburg, “From persistent identifiers to digital objects to make data science more efficient”, *Data Intelligence*, 1 (2019): 6-21. DOI: http://doi.org/10.1162/dint_a_00004.

²² E. Plomp, “Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments”, *Data Science Journal*, 19(1) (2020):, 46. DOI: <http://doi.org/10.5334/dsj-2020-046>.

Digital Object Identifier (DOI)

Digital Object Identifier is the most widely adopted PID for different kinds of research objects and publications. DOI is a unique, persistent digital identifier of an object— digital or physical used extensively in scholarly publishing. DOI promotes discovery and interlinking and can be assigned to a preprint, journal, journal article, book, book chapter, conference proceeding paper, dataset, presentation, image, table, etc. DOI provides a persistent link to an object and standard metadata for that object.

The DOI system is managed by the International DOI Foundation (IDF), a not-for-profit membership organisation that is the governance and management body for the federation of Registration Agencies providing DOI services and registration and is the registration authority for the ISO standard (ISO 26324) for the DOI system²³.

The DOI system implements the Handle System²⁴ (a general-purpose global name service enabling secure name resolution over the Internet) and the index Framework²⁵ (a generic ontology-based contextual data model structure). According to the DOI Handbook²⁶, the DOI system provides a specified standard numbering syntax, a resolution service, a data model incorporating a data dictionary, and an implementation mechanism through a social infrastructure of organisations, policies and procedures for the governance and registration of DOI names.

DOI name syntax consists of prefix/suffix, e.g. 10.1016/159, where “10” is the DOI identifier within the Handle system, “1016” is a registrant code of the organisation that has assigned the DOI, and the suffix “159” identifies the resource, separated by a “/”. Each suffix shall be unique to the prefix element that precedes it. DOI is usually expressed on the Web as URL: <http://dx.doi.org/10.1016/159>. The prefix and suffix can be subdivided further since DOI is an opaque string with no embedded meaning and limits on the length. A DOI name may be assigned to any entity, which must be precisely defined through structured metadata. The DOI name itself remains persistent through ownership changes and unaltered once assigned.

Noteworthy among the group of DOI Registration Agencies are Crossref²⁷ and DataCite²⁸. Crossref is an organisation registered in the United States in 2000 and run by the Publishers International Linking Association (PILA). Today Crossref is working with 17.000 members from 140+ countries, agreeing to assign DOIs to their current content (130+ million records). Crossref assigning DOIs for journal articles, books and book chapters, conference proceedings and conference papers, technical reports and working papers, theses and dissertations, peer reviews, grants, preprints, standards, databases and datasets and components. DataCite was founded in the UK in 2009 to improve data citation, establish easier access to research data, increase acceptance of research data and support data archiving.

²³ <https://www.doi.org>

²⁴ <https://www.handle.net>

²⁵ https://www.doi.org/factsheets/indecs_factsheet.html

²⁶ https://www.doi.org/doi_handbook

²⁷ <https://www.crossref.org>

²⁸ <https://datacite.org>

3.1.2. Open Researcher and Contributor ID (ORCID)

Problems with one name assigned to multiple persons²⁹, multiple names assigned to one person, overuse of the initials and abbreviations instead of full names which stay unresolved, pseudonyms, missing names/surnames, misspelt names, added names, merged names, ordering of given names and surnames and changed names have been present in scholarly publishing for centuries. Finally, inconsistent journal practices and inappropriate cataloguing rules heighten the confusion.

Different attempts were made to solve these problems, but only in limited environments. Some bibliographic databases employ their own identifiers like ResearcherID (Web of Science Core Collection, Clarivate) or Scopus Author ID (Scopus, Elsevier), as are some repositories and archives (e.g. arXivID). Researchers could have an identifier which is unique inside one country. Still, such solutions are not open and globally recognisable.

A globally accepted and unique Google Scholar ID identifier has become very popular across the academic community, improving researchers' visibility. Still, we should not forget that Google is a commercial company. Google Scholar will remain available for as long as they believe it to be a successful, or at the very least, not overly expensive, component of their business strategy. Therefore, Google Scholar ID could not be considered a persistent identifier.

ORCID (Open Researcher and Contributor ID) is an open-source, cross-national identification system that provides a persistent digital identifier (an ORCID ID) that a researcher owns and controls, distinguishing her/him from every other researcher³⁰. ORCID gives people a unique identity when engaging in research, scholarship, and innovation activities. Its goal is to enable transparent and reliable links between researchers, their publications, peer review and other contributions, grants and affiliations. Researchers may include their ORCID identifier when they write a data management plan, deposit a dataset into a repository, or access a dataset for analysis purposes³¹. Since its founding in 2012, ORCID has benefited the research community by making it possible for persistent identifiers and metadata to be collected, connected, and reused under the complete authority of the researchers who use them³². ORCID provides a free, non-proprietary registry of persistent unique identifiers for researchers, scholars, and analysts, together with APIs that enable the interoperable exchange of information between systems to embed identifiers in research systems and workflows.

ORCID ID profiles can be connected with unique identifiers assigned by different services like Google Scholar ID (Google), ResearcherID (Clarivate), Scopus Author ID (Elsevier), etc. That does not mean that these connections between different PIDs for the same person are explicitly stated by different stakeholders, hence the need to harmonise, deduplicate and normalise this type of data.

²⁹ Estimates by China's Ministry of Public Security suggest that more than 1.1 billion people — around 85% of China's population — share just 129 surnames.

³⁰ <https://orcid.org/>

³¹ L. L. Haak, 'Persistent Identifiers Can Improve Provenance and Attribution and Encourage Sharing of Research Results'. 1 Jan. 2014: 93 – 96. DOI: 10.3233/ISU-140736.

³² Shillum, Chris; Petro, Julie Anne; Demeranville, Tom; Wijnbergen, Ivo; Hershberger, Sarah; Simpson, Will (2021): From Vision to Value: ORCID's 2022–2025 Strategic Plan. ORCID. Online resource. <https://doi.org/10.23640/07243.16687207.v1>.

3.1.3 Research Organization Registry (ROR)

ROR is a community-led project developing an open, sustainable, usable, and unique identifier for every research organisation in the world by providing identifiers that are globally unique, stable, discoverable, and resolvable. In addition, ROR develops appropriate metadata schema for organisations and explores interoperability with other identifiers through relationship metadata³³. ROR is intended for use by the research community to increase the use of organisation identifiers and enable connections between organisation records in various systems.

Starting with a group of 17 organisations, ROR is intended for use by the research community to increase the use of organisation identifiers in the community and enable connections between organisation records in various systems.

Access to organisations for managing ROR records shall be via permission. ROR focuses on the organisation levels most pertinent for the affiliation use case (who employs, educates, funds, etc.). According to ROR documentation, required metadata, and Open Definition conformant license, the organisation should provide metadata elements sufficient to uniquely identify the organisation in human- and machine-readable formats. ROR also provides open criteria and documented processes for inclusion/exclusion, creating, merging, and deprecating an institution's record. Record changes are tracked and recorded using an open provenance model. ROR also claims a robust customer support system and an open knowledge base to maintain a good relationship with the community's technical teams.

Other relevant PIDs enable linking of different aspects of SSH research, like Archival Resource Key (ARK)³⁴, Dewey³⁵, Entertainment Identifier Registry Association (EIDR)³⁶, VIAF³⁷, Open Funder Registry (FundRef)³⁸, Contributor Roles Taxonomy (CredLiT)³⁹ developing a taxonomy for contributors to research output, Persistent Identification of Instruments (PIDINST)⁴⁰, International Geo Sample Number (IGSN)⁴¹, Research Resource Identifiers (RRID)⁴², and Research Activity Identifier (RAID)⁴³.

Challenges for the humanities

Although using PIDs has solved many problems related to finding digital sources, there are many challenges in their wider adoption, especially in the SSH and the humanities specifically.

First of all, humanities data are closely related to cultural heritage data. For example, many of the documents humanities researchers are analysing are originating from GLAM collections (i.e. archival documents, librarian catalogues, literary works, musical notations

³³ <https://ror.org/>

³⁴ <https://arks.org/about/>

³⁵ <https://www.gutenberg.org/files/12513/12513-h/12513-h.htm/>

³⁶ <https://www.eidr.org/about-us/>

³⁷ <https://www.oclc.org/en/viaf.html/>

³⁸ https://gitlab.com/crossref/open_funder_registry/

³⁹ <https://credit.niso.org/>

⁴⁰ <https://www.rd-alliance.org/groups/persistent-identification-instruments-wg/>

⁴¹ <https://www.igsn.org/>

⁴² <https://www.rrids.org/>

⁴³ <https://www.raid.org.au/>

etc.), historically it is not always clear who could be considered a researcher or a creator, finally many institutions in the humanities have hybrid profile (i.e. publishers publishing cultural books and research monographs). This impacts the set of PIDs which are relevant for the humanities. A good selection of humanities-specific PIDs can be found in the *Developing Identifiers for Heritage Collections*⁴⁴.

Second, there is a specificity to the humanities publishing environment. For humanities stakeholders it is important and necessary to cherish the bibliodiversity⁴⁵ of the output. For example, books⁴⁶ play a large role in the humanities – relatively larger than in other disciplines – which pose a set of specific issues for the humanities. One of them is the issue of granularity – should PIDs be attributed at the level of a whole book, or the chapters (and how to define all complex variations of this challenge). Another type of humanities output that adds to the diversity are cultural heritage publications, especially originating from smaller publishers, such as cultural magazines or newspapers, or citizen science publications (i.e. blogs). These publishers or actors on many occasions do not have access to digital infrastructures providing PIDs or do not possess relevant know-how to address this issue. Finally, the historical output is even more relevant to humanities research, than in other sciences where the most current findings seem to circulate more dynamically. This is exemplified by humanities’ publications referencing older publications, authors and non-digitised content (and for these entities PID attribution is not equally incentivised).

Third, humanities research output is multilingual⁴⁷. For certain types of PIDs – like thesauri, controlled vocabularies, taxonomies etc. – this poses a serious challenge of making the resource understandable in local languages. For example, Library of Congress subject headings are mostly accessible in large European languages (German, French), but not smaller ones (like Czech, Polish or Croatian). The lack of effort to provide these kinds of resources in a truly multilingual fashion will limit the uptake of PIDs in smaller countries and in less widely-used languages.

Fourth, some challenges are related to the creation of PIDs and the costs, which especially influence the humanities and their set of smaller stakeholders which are not always incentivised or have the capacity to implement PID systems.. An individual researcher can obtain an ORCID ID and have control over its content and what information will be publicly available and displayed. Still, the institution must pay for institutional membership to access enhanced services. However, the process of obtaining a DOI is different since DOI must be registered via a registration agency which needs to be a member of Crossref, DataCite or another institution. Membership fees are not cheap, especially for small academic institutions and learned societies. Furthermore, each assigned DOI is charged separately. Together with DOI policy requirements recommending which types of objects can be associated with DOI, the additional fees lead to “savings” on permanent identifiers, even for the institutions and journals that can afford annual membership. To become more aligned with equity, diversity and inclusion principles of open science, PID providers’ pricing should be tiered more appropriately. For example, a ROR ID could be created free of charge.

⁴⁴ R. Kotarski et al., *Developing Identifiers for Heritage Collections*. Zenodo. 2021. DOI: <https://doi.org/10.5281/zenodo.5205757>.

⁴⁵K. Shearer, L. Chan, I. Kuchma, P. Mounier, *Fostering Bibliodiversity in Scholarly Communications: A Call for Action*. Zenodo. 2020, DOI: <https://doi.org/10.5281/zenodo.3752923>.

⁴⁶Ferwerda, Eelco, Pinter, Frances, & Stern, Niels. *A Landscape Study on Open Access and Monographs: Policies, Funding and Publishing in Eight European Countries*. Zenodo. 2017. DOI: <https://doi.org/10.5281/zenodo.815932>.

⁴⁷ D. Leão, M. Angelaki, A. Bertino, S. Dumouchel, & F. Vidal. *OPERAS Multilingualism White Paper*. Zenodo. 2018. DOI: <https://doi.org/10.5281/zenodo.1324026>.

Besides the specific humanities context, there are still uncertainties and challenges which impact the whole PID ecosystem (including the humanities). Existing security, reliability and resilience issues related to the unauthorised changes of PID databases and insufficient maintenance strategies were recorded. Accordingly, although persistency is presumed when considering PIDs use, some studies show that persistency is not warranted and that scholarly content providers respond differently to varying request methods and network environments and even change their response to requests against the same DOI⁴⁸. Furthermore, there is a need for sustainable services also for non-data resources, a global resolution service for any kind of PID, and a common mechanism for complex querying across PID systems⁴⁹.

Some groups of experts are already working on establishing data infrastructure based on digital object access protocol, which represents a universal exchange protocol for digital objects stored in repositories using different data models and organisations⁵⁰. According to Wittenburg, 2019 this approach can solve some fundamental problems in data management and processing.

Conclusion

The uptake of PIDs for publications, data, software, researchers and research organisations has increased in recent years. Still, the widespread adoption of PIDs relevant to other aspects of research has yet to be realised.

If we look at the specificity of the disciplines in the field of SSH, it is clear that it will be challenging to establish standards that can be easily applied within all disciplines and sub-disciplines. It is to be assumed that common PIDs will be extended with discipline-specific standards, especially with persistent identifiers for physical samples, artefacts, reagents and instruments⁵¹. Therefore, some PIDs will have to adapt their metadata and categorisation schemas, registries, controlled vocabularies and ontologies to support a broader diversity of research from different disciplines.

⁴⁸ M. Klein, L. Balakireva, "On the persistence of persistent identifiers of the Scholarly Web", 2020. DOI: <https://arxiv.org/abs/2004.03011>.

⁴⁹ European Commission (2020): PID Architecture for the European Open Science Cloud. DOI: <https://doi.org/10.2777/525581>.

⁵⁰ P. Wittenburg, "From persistent identifiers to digital objects to make data science more efficient", *Data Intelligence*, 1 (2019): 6-21. DOI: http://doi.org/10.1162/dint_a_00004.

⁵¹ E. Plomp, "Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments", *Data Science Journal*, 19(1) (2020):46. DOI: <http://doi.org/10.5334/dsj-2020-046>.

3.2. GoTriple, Dublin Core and Persistent Identifiers

Cezary Rosiński⁵², Tomasz Umerle⁵³, Nikodem Wołczuk⁵⁴

3.2.1 PIDs in GoTriple

The Data model for GoTriple is formalised by using Schema.Org as a base ontology. However, the data extraction relies on integration with external data models and formats, such as the Isidore research discovery platform dedicated from Fren outputs from the Social Sciences and Humanities, OpenAIRE, Europeana Data Model and Dublin Core (DC). As a large part of the current data is harvested by OAI-PMH/DC sources and metadata issues interesting for SSH aggregation are particularly visible in this format, in this chapter we will focus solely on problems with persistent identifiers in DC format delivered through OAI-PMH. An OAI-PMH implementation must be able to display metadata in Dublin Core format, however, it may also have the capability to support additional formats. Therefore, the default format for the OAI-PMH protocol is OAI-DC – also very common among data providers (e.g. Directory of Open Access Journals⁵⁵, Biblioteka Nauk [Library of Science]⁵⁶, EKT⁵⁷) easily readable by both humans and XML parsing tools, but very limited in terms of the ability to properly represent complex data relations. DC facilitates aggregation of a large number of sources. Yet, this choice has its consequences for the PID attribution in GoTriple, as DC is a “flat” format which makes it difficult to handle relationships between values and external identifiers⁵⁸.

It is not necessarily the case that GoTriple data providers do not offer PID-enriched data, but it is common that the DC-based expression of their data is not equipped with PIDs. One example is the Polish Bibliotekanauki.pl where records are exposed in three formats: DC, JATS and BWMETA. Of the referenced formats, only JATS (not DC!) contains additional identifiers like ORCID and ISSN separately and inside the intended field. It is noteworthy that the ISSN identifier does exist in DC records but it is placed in the tags conveying also different types of information (such as “source” where name of journal is placed or “identifier” where DOI can be stored). The statistic below shows the coverage of some of the data identifiers from bibliotekanauki.pl in the JATS format.

⁵² Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-6136-7186>.

⁵³ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-7335-0568>.

⁵⁴ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-4303-2016>.

⁵⁵ <https://doaj.org/>

⁵⁶ <https://bibliotekanauki.pl/>

⁵⁷ <https://www.ekt.gr/en/index>

⁵⁸ D. M. Vogel, “Qualified Dublin Core and the Scholarly Works Application Profile: A Practical Comparison”, *Library Philosophy and Practice (e-journal)*, 1085 (2014). <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=2685&context=libphilprac>

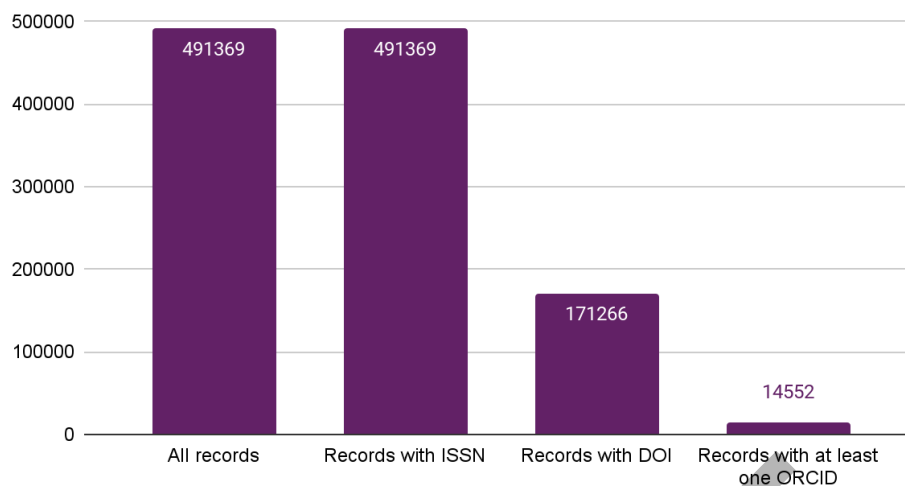


Figure 3.2.1. PIDs from Library of Science in JATS format (*bibliotekanauki.pl*, September 2022).

We will focus, in this text, on the possibilities of PID enrichment for GoTriple data, not on the discussion of which PIDs should be chosen and implemented for this enrichment. For this discussion we will, on the one hand, present the current development of the DC format and on the other present possible broader strategies for PID enrichment, as the DC-based approach will not be sufficient in the near future.

3.2.2 PIDs in Dublin Core – will DC ever be good enough?

DC provides solutions for preserving PIDs in the tag `dc:identifier`, which contains mostly a document identifier such as DOI. The other PIDs presented in DC are usually ISSN for journals and ISBN for books. However, that leads to two kinds of issues. First all of the identifiers for different types of data (e.g. document vs. journal) are described with the same multiplied tag. Second, in cases of multiple entities of the same type, it prevents identifying them by connecting the phrase to the ID.

Dublin Core Metadata Initiative (DCMI) – especially the PIDs in Dublin Core Working Group⁵⁹ – is conducting work to analyse existing practices and opportunities related to the presentation of identifiers. The aim is to develop recommendations on how to correctly represent identifiers with their corresponding strings (the problem concerns, for example, information about the author of the publication) using DC format tags and attributes. Various approaches are possible, of which DCMI considers two.

⁵⁹ <https://www.dublincore.org/groups/pids-in-dc-wg/>

1. To use an attribute (`id=`) to hold a PID for an XML element with a literal value:

```
<dc:creator id="https://orcid.org/0000-0003-1541-5631">Walk, Paul</dc:creator>
```

2. To use one or more DC Identifier elements in a nested description of the entity in question:

```
<dc:creator>
  <dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
  <dc:identifier>http://paulwalk.net</dc:identifier>
  <foaf:name>Walk, Paul</foaf:name>
</dc:creator>
```

Walk, Paul. (2019, January 28). PIDs in Dublin Core.

Figure 3.2.2. Approaches to store PIDs in DC considered by DCMI (Walk, Paul. 2019, January 28. PIDs in Dublin Core. Zenodo. <https://doi.org/10.5281/zenodo.2551181>).

One of the suggested solutions presented above equips the DC piece of information with additional structural elements such as an attribute. The attribute contains key-value construction where external identifiers may be presented as a property. The other solution extends the somehow flat structure by providing the nested properties for identifiers, where additional levels of the XML tree may be built.

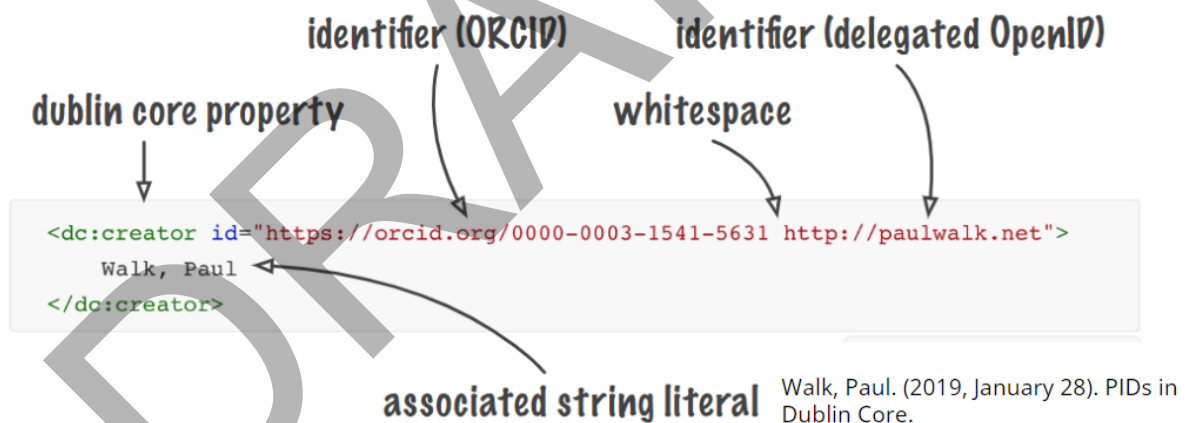


Figure 3.2.4. Using tag attributes to store PIDs (Walk, Paul. 2019, January 28. PIDs in Dublin Core. Zenodo. <https://doi.org/10.5281/zenodo.2551181>).

The proposed solutions solve the problems mentioned earlier in the representation of identifiers, but would still require extending the simplified format used by the OAI-PMH protocols, which is not a simple task. This requires interference with the protocol and the initiative of providers to adjust the data provided so far.

However, as long as the main data format used in GoTriple to aggregate data is the simplified OAI-DC – quite indifferent to the DC development – the problem seems impossible to solve. The main goal of the platform in this area should be to move toward more flexible data exchange formats that will allow for the appropriate extraction and storage of identifier data, such as JATS and BITS.

3.2.3 How to propagate PIDs in GoTriple in the future?

Once the data format used in GoTriple has been adapted to properly store and represent persistent identifiers and exchange protocols would be flexible enough to handle additional data, it is important to take future steps to enrich the data with missing identifiers. It seems that GoTriple can apply three different strategies to enrich the aggregated data, but each comes with additional challenges.

- Different formats

The first strategy is to use other, often richer formats offered by the provider. An example is the Polish provider Biblioteka Nauki, which presents much richer data for articles in JATS format than in DC format. The GoTriple platform already started processing data in additional formats and consistently expands the formats diversity (EDM for OAI-PMH, soon the DC extended by BASE named BASE-DC* and for data dumps the ISIDORE and OpenAIRE formats). While it is important to significantly expand the range of data acquired, one should not forget that this can only come at the cost of adapting the service to support a larger number of differing formats. Implementing support for such a large variability requires in fact extensive substantive and technical work.

- Disperse data and providers/aggregators with richer data

The first solution does not address the issue of data dispersion among several data sources, i.e. the same records (e.g. publications) may be stored in many places and be described differently by various databases. For the sake of state-of-the-art aggregation of SSH data – which is the aim of GoTriple – a preprocessing mechanism should be constantly supported to identify the same content presented in multiple sources, so as to obtain the biggest possible PID coverage. This facilitates a number of other data processing-related tasks, such as normalisation (e.g. deduplication), enrichment, and harmonisation of the data. It is easier to handle the general heterogeneity of GoTriple contents with PID-rich metadata for documents (even though the PIDs are coming from different resources).

- Using internal identifiers

The last strategy is to use one's own persistent identifiers. While this approach avoids the problems associated with identifying resources elsewhere and acquiring data in new formats, it raises a number of other challenges. Assigning internal identifiers requires their subsequent continuous control and maintenance of additional workflow. An approach where a new data schema is created instead of using already well-established and widely used solutions can also be considered inappropriate. However, the use of internal identifiers is only useful in the closed ecosystem of the service. It is of little use deviating from current Linked Open Data practices until other users adapt it for their own use.

These solutions seem to be a good way of thinking about data enrichment, but they require serious action on both the providers' and GoTriple platform's sides. Trying to implement the new solutions with multiple providers at the same time can be a major challenge. A large dataset like GoTriple demands preprocessing, tracking of different sources due to the current data ecosystem. It is necessary to at least monitor and assess the quality in multiple providers to properly adapt to the changes. Appropriately, the TRIPLE consortium is both securing the sustainability of current data processing solutions and IT work on maintaining the GoTriple, but also actively looks for future development that will allow for constant improvement and flexibility of data workflows.

DRAFT

3.3.Guidelines for PIDs as humanities research data in the context of TRIPLE – summary

1. Machine-readable PIDs for people, organisations or publications are widely accepted and continuously implemented throughout the scholarly data ecosystem. However, the PID should be extended and adapted to the discipline-specific demands. For the humanities it means, especially, **considering application of the PIDs outside of the scientific ecosystem** – such as PIDs used in the cultural heritage domain which are common in the cultural heritage sector (e.g. in libraries). As currently PIDs are not only identifiers, but also connectors, **enabling linking between different resources**, it is important to leverage PID's potential to connect scientific and cultural output which is crucial for the humanities.
2. Specific needs of the humanities stem also from the importance of **bibliodiversity and multilingualism** in humanities which calls for humanities-specific actions and projects that will properly adapt PID standards to different publication genres and/or local contexts. Additionally, the diversity of the actors involved in the humanities ecosystem – i.e. cultural institutions, NGOs etc. – might make the PIDs' uptake more challenging (due to the lack of resources, know-how of cross-sectoral interoperability).
3. PID coverage grows also in the humanities which are more dependent on smaller data providers. This does not mean that this data is present in a way that the aggregator mechanism finds the most efficient, sustainable or proper. To build a PID-rich and reliable humanities research dataset through metadata aggregation – which is the goal of GoTriple as long as it aims to extract any knowledge from aggregated resources – a **complex approach is needed that either takes advantage of the multiplicity of providers' endpoints or is able to enrich data with external identifiers**.
4. Dispersion of documents, scattered among multiple providers, demands deduplication that is also sensitive to the fact that relevant PIDs can be provided by only one (some) of the providers. **Building large humanities research dataset based on publications metadata demands a level of analysis of multiple data sources**. This analysis – preferably a dynamic and continuous one – is a prerequisite for fuller enrichment of a dataset.

4. Keywords

4.1. Keywords in the SSH – current state and upcoming challenges

Cezary Rosiński⁶⁰

4.1.1 Definition and application of keywords

Keywords are a central issue not only in scientific production, but also in the storage of the results of researchers' work. On the one hand, they are regarded as meaningful words taken from the title or text of a document to represent its content and are used in most scholarly articles to describe or summarise their content. The most common source of keywords is author, indexer (e.g. humans or machines who provide metadata enrichment to scholarly indexing services such as bibliographers or library catalogue systems) or both. On the other, and just as importantly, keywords capture the essence of the topic, make it easier to classify content and lead users to resources. In this sense, keywords are a tool for communication between researchers and communities interested in the results of their work and therefore, they play a crucial role in research discovery systems. In practice, however – due to the number of articles and books published – sets of keywords that describe a specific resource but are not drawn from a controlled vocabulary get lost among other metadata. Keywords are then just a collection of unrelated and unstructured phrases that do not lend themselves to re-use.

Keywords serve two purposes: describing and cataloguing resources.⁶¹ These are, to a degree, contradictory use cases. Describing means characterising a document with proper terminology (resulting in highly detailed headings that correspond as closely as possible to the content of the document) while cataloguing means helping to find the document in the database (accepting mechanisms for aggregating information and applying a degree of generalisation to provide search terms that apply across publications with similar content). Both use cases require specific competencies. How to make authors familiar with classifications (when they are researchers, not catalogers working with a specific database) vs. how to make indexers more competent in different fields of research (in fact, very detailed research)?

4.1.2 Typology of keywords

⁶⁰ IBL PAN, <https://orcid.org/0000-0002-6136-7186>.

⁶¹ C. Rockelle Strader, "Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations", *Library Resources & Technical Services* 53 4 (2009): 250.

Understanding the role of keywords in the information search process requires consideration of their form and origin in addition to their functionality. An initial typology of keywords is based on their form, including **strings/lists of strings**, **subject headings** and **controlled vocabularies**. It is best to consider these three types as stages in the development of knowledge presentation and search mechanisms.

String keywords are a collection of phrases lifted directly from the text that are not linked to any external information source. This means that not only do they not allow for any search, but they cannot be used to filter content or find content that is semantically similar but expressed in a different form. Subject headings offer slightly more possibilities. All occurrences of the same keyword are linked, making it easier to filter content and get search results that match certain parameters and they are selected from a list of subject headings that includes preferred forms of terms. Subject headings are also used in bibliographies and indexes, acting as an access point and allowing you to search for a work by subject in a library/bibliographic cataloguing database. The most advanced environment for keywords is a controlled vocabulary. It is a structured collection of words and phrases (subject headings, nomenclature of persons and institutions) and their relations to each other used for content indexing and retrieval. It includes preferred as well as non-preferred terms, where preferred terms are used for indexing and has a specific scope or describes a specific domain. The use of a controlled vocabulary improves search results, and its implementation organises information and gives access to structured system resources.

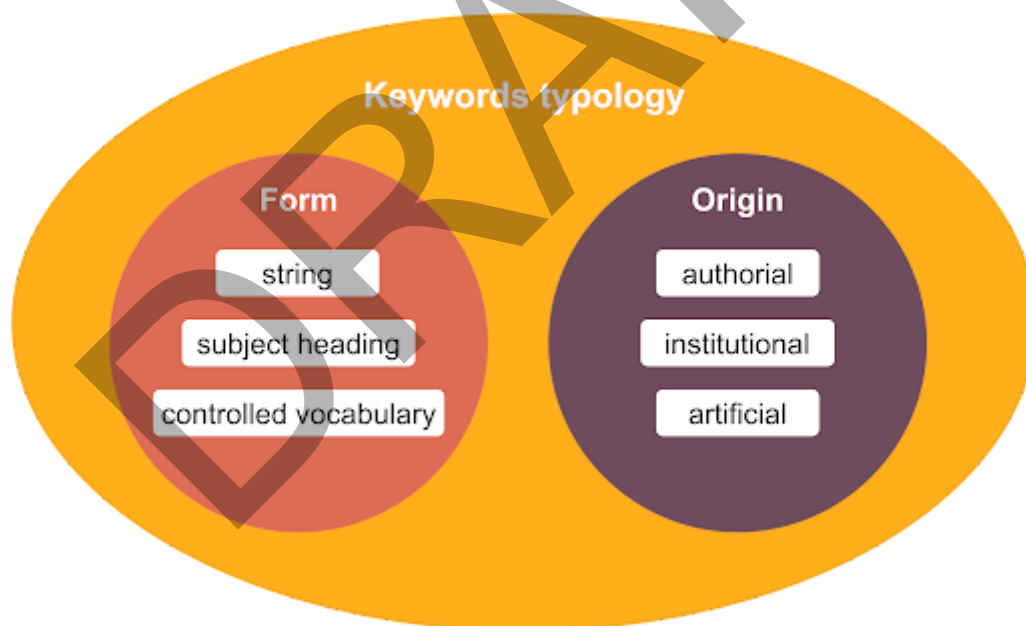


Figure 4.1.1. Keywords typology

The alternative typology of keywords includes information on their source and includes **authorial**, **institutional** and **artificial origins**. The author keywords are assigned by scholars who are subject specialists, which provides the highest degree of descriptive detail. At the same time, the authorial keywords display the highest sensitivity to nuances, as authors are alert to any changes or shifts in terminology. In addition, they are characterised by a focus on the most relevant subjects, and no restriction on the type or number of keywords. Author generated keywords can come from a controlled vocabulary (for instance,

selecting discipline from a drop-down while depositing content in a repository) or, much more commonly, from free text.

Institutional keywords are often assigned by general bibliographers rather than field-specific bibliographers, thus shifting the focus from the best way of describing a text to the best way to document it within its given database in order to maximise information retrieval. A vocabulary is oriented towards cataloguing, therefore it is also focused on looking for similarities. The need to fit into a database framework means institutional keywords must first and foremost adhere to an existing structure, making it more difficult to capture nuance and change, while making it easier to capture similarities to existing data.

The last – and newest – way of creating keywords is using artificial intelligence⁶². This issue will be exhaustively discussed in the section "Keywords generation in SSH", while here it will only be outlined. Artificial keywords are the most resource efficient way of describing content, as they are generated on a massive scale and can handle the ever-increasing number of scientific texts. However, it should be noted that they introduce additional problems into the keyword creation environment: they are difficult to verify, are limited by language models, and often pose the problem of balancing precision and recall. Keywords created in this way usually appear as unrelated strings and with no link to controlled vocabularies. An example of automatic tools used in cataloguing is the Finnish software Annif⁶³, a tool for automated subject indexing and classification. Annif employs a combination of existing natural language processing and machine learning tools including TensorFlow, Omikuji, fastText and Gensim. It is multilingual and can support any subject vocabulary (in SKOS or a simple TSV format). It provides a command-line interface, a simple Web UI and a microservice-style REST API.

4.1.3 Keywords and Linked Open Data

Linked Open Data⁶⁴ (LOD) – according to the World Wide Web Consortium definition – is a vision of globally accessible and linked data on the internet based on the Resource Description Framework (RDF) standards⁶⁵ of the semantic web. LOD is often thought of as a virtual data cloud where anyone can access any data they are authorised to see and may also add to any data without disturbing the original data source. This provides an open environment where data can be created, connected and consumed on an internet scale. A basic theory of LOD is that data has more value if it can be connected to other data. Data, in this context, is any structured web-based information. LOD has been proposed as the basis for open data governance and for solving many of the data integration issues. LOD helps build bridges between different formats and allows connecting different interoperable information sources. This can be achieved by linking the resource to an entity of a well-known value (e.g. a Wikidata/Wikipedia URI). As a result, data integration and viewing complex data become easier and more efficient.

⁶² K. Golub, "Evaluating automatic subject indexing: A framework", Keynote speech at the 7th ISKO Italy Meeting Bologna, 20 April 2015. <http://www.iskoi.org/doc/bologna15/golub.htm>

⁶³ O. Suominen, J. Inkinen, M. Lehtinen. "Annif and Finto AI: Developing and Implementing Automated Subject Indexing". *JLIS.it* 13, 1 (2022): 265-82. <https://www.jlis.it/index.php/jlis/article/view/437>

⁶⁴ F. Bauer, M. Kaltenböck, "Linked Open Data: The Essentials", Vienna: edition mono/monochrom, 2012. https://cdn.semantic-web.com/wp-content/uploads/2017/05/LOD-the-Essentials_0.pdf.

⁶⁵ <https://www.w3.org/RDF/>

4.2. Keyword generation in SSH

Agnieszka Mikołajczyk-Bareła⁶⁶, Agnieszka Karlińska⁶⁷, Maciej Ogrodniczuk⁶⁸, Piotr Pęzik⁶⁹

Introduction

Keywords used as descriptors/metadata of scholarly texts are usually nouns or longer nominal phrases, which succinctly describe the article's content. Such keywords can be abstractive in that they do not necessarily need to occur verbatim in the running text of a document to be considered significant. An important aspect in keyword generation is the nature of the vocabulary: it can be either controlled or uncontrolled.

4.2.1 GoTriple's wishlist

What would be the perfect keyword generation algorithm from the GoTriple users perspective?

- A solution that assigns keywords to articles, based either on abstracts or, if available, the full text of a paper;
- A solution that generates conventional generalisations of content, e.g. “postmodernism”, “literary theory”, but at the same time extracts words and phrases specific to the paper – keywords can be identified extractively, i.e., based on literal occurrences in the text, or abstractively, i.e., are a generalised description of it;
- These can be single or multi-word phrases, as long as they are complete syntactically/semantically;
- Noun phrases are preferred;
- Keywords should be lemmatized and in the case of longer phrases, the syntactic agreement between phrase components should be preserved or recovered. Moreover, the proper casing of named entities should be used, cf. “Tatra National Park”, “Institute of National Remembrance”;
- Relevance of keyword assignments is preferred over coverage (a set of precise but incomplete keywords is generally better than a set of complete but imprecise keywords);
- For abstracts the number of keywords identified should average between 3 and 5 items. For full-text, the number of keywords returned should be determined by the user;
- The keyword assignment method should be able to yield keywords from both an open, bottom-up, uncontrolled vocabulary and a controlled vocabulary;
- It should perform well for many languages, both resource-rich and low-resource, and many scientific disciplines.

⁶⁶ Voicelab, <https://orcid.org/0000-0002-8003-6243>.

⁶⁷ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-4846-7086>.

⁶⁸ Institute of Computer Science, Polish Academy of Sciences, <https://orcid.org/0000-0002-3467-9424>.

⁶⁹ University of Lodz, <https://orcid.org/0000-0003-0019-5840>.

4.2.2 Keywords automated generation methods

Early works regarding automated keywords extraction and generation goes back to the 1990s. One of the first approaches used decision trees to assign each phrase from the paper a binary label: a keyphrase or not a keyphrase. Since then, many other approaches emerged, such as treating keyword extraction as, for instance, a statistical task, an extreme multi-label text classification, or even text generation. Statistical methods like TfIdf or KP-Miner⁷⁰, analyse the keyphrase frequencies, position, and sometimes context to find the most relevant ones. Recent methods based on more or less “deep” machine learning models treat keyphrase extraction often as an extreme classification problem⁷¹: they assign each section of text to a few classes (keywords). This results in thousands, or even millions, of possible classes (here – keyphrases) that each text can be assigned to. Since it is a multi-labeling problem (each input can be assigned to multiple classes), it results in a very complicated, “extreme” problem. Finally, most current studies focus on text generation⁷². They use large language generation models pretrained on terabytes of data, and train them to extract and generate keyphrases based on the input text. In this section we will present, train and test a few commonly used methods of keyword extraction based on scientific articles corpora called CURLICAT⁷³. For a more in-depth review of keyword assignment methods see Papagiannopoulou and Tsoumakas⁷⁴.

4.2.3 Controlled vocabulary

Keyword extraction

A brief overview of keyword extraction tools is worth starting with ExtremeText,⁷⁵ which is an extension of a popular text classification library called FastText⁷⁶. FastText uses vector representation of subwords to train relatively shallow neural networks. ExtremeText builds on that, by using Hierarchical softmax classifier. It uses both Probabilistic Labels Tree (PLT) loss and k-means clustering for hierarchical tree building. This allows for training models on very large taxonomies with hundreds of thousands classes. The downside is that it might

⁷⁰ S. R. El-Beltagy, A. Rafea, “KP-Miner: Participation in SemEval-2”, Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala 2010: 190-193, <https://aclanthology.org/S10-1041.pdf>

⁷¹ W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. D. Dhillon, “Taming Pretrained Transformers for Extreme Multi-label Text Classification”, KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York 2020: 3163-3171, https://dl.acm.org/doi/abs/10.1145/3394486.3403368?casa_token=-E8vT-FHdnwAAAAA:PmQdjkC9gsvvgsZok4T3MptmbPNOzh6DyGC1MkElwOwDnwaV1S0OF5IXKsRVVol7tqgBOjG3Xmc.

⁷² P. Pęzik, A. Mikołajczyk, A. Wawrzyński, B. Nitoń, M. Ogrodniczuk, “Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer”, ACIIDS2022, <https://arxiv.org/abs/2209.14008>.

⁷³ P. Pęzik, A. Mikołajczyk, A. Wawrzyński, B. Nitoń, M. Ogrodniczuk, “Keyword...”.

⁷⁴ E. Papagiannopoulou, G. Tsoumakas, “A Review of Keyphrase Extraction”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10 (2), <https://arxiv.org/pdf/1905.05044.pdf>.

⁷⁵ M. Wydmuch, K. Jasinska, M. Kuznetsov, R. Busa-Fekete, K. Dembczyński, “A no-regret generalization of hierarchical softmax to extreme multi-label classification”, Advances in Neural Information Processing Systems 31, 2018.

⁷⁶ A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, “Bag of Tricks for Efficient Text Classification”, In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain 2017: 427–431.

have low portability to out-of-distribution domains that were not included in the dataset. Additionally, the vocabulary is controlled, hence it poses challenges for rapidly growing disciplines where new terminology appears every year or even every month.

4.2.4 Uncontrolled vocabulary

Keywords extraction

KeyBERT⁷⁷ is yet another popular approach to keyword extraction, this time instead of extreme classification focusing on the unsupervised generation of keywords. The methods adapt the BERT transformer model⁷⁸ by creating vector representations of n-grams and comparing them to the vector representation of a whole document. The ranking of key terms is calculated according to cosine similarity between n-gram and document vectors. The method is purely 'extractive', as terms are straightforwardly copied and pasted from the text. It is characterised by an arbitrary operationalization of key terms, only n-grams, with no normalisation. The advantage is that there is no need to train the model as it can use pretrained BERT available online. The fine-tuning for keyword extraction would give a user an advantage.

Another keyword extraction tool worth taking into account is TermoPL⁷⁹. It is an algorithm designed for terminology extraction from corpora. It identifies terms that match syntactic patterns and produces a ranking based on, among other things, a variant of the C/NC-value measure⁸⁰. As it identifies, lemmatizes, and scores recurrent noun phrases as potential terms using a ranking function, it can be used for keywords extraction. Like KeyBERT it is a pure extraction approach.

Keywords generation

One of the latest keyword generation models is vIT5, which is based on encoder-decoder architecture using Transformer blocks presented by Google⁸¹. The input to the model is text preceded by a prefix, and the output is the target text, where the prefix defines the type of task: e.g. "Translate from Polish to English:". The vIT5 was trained on a scientific articles corpus to predict a given set of keyphrases based on the concatenation of the article's abstract and title. It generates precise, yet not always complete keyphrases that describe the content of the article based only on the abstract. The biggest advantage is the transferability of the vIT5 model, as it works well on all domains and types of text. The downside is that the text length, as well as the number of keywords, is similar to the training data: the text piece of an abstract length generates approximately 3 to 5 keywords. It works both extractive and

⁷⁷ M. Grootendorst, *Keybert: Minimal Keyword Extraction with Bert*, 2020, [10.5281/zenodo.4461265](https://zenodo.org/record/4461265).

⁷⁸ J. Devlin, M-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *NAACL-HLT* (1) 2019: 4171–4186.

⁷⁹ M. Marciniak, A. Mykowiecka, and P. Rychlik, "TermoPL — a flexible tool for terminology extraction". In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia: 2278–2284*, European Language Resources Association (ELRA), European Language Resources Association (ELRA).

⁸⁰ K. Frantzi, S. Ananiadou, H. Mima, "Automatic Recognition of Multi-Word Terms: the Cvalue/NC-value Method". *Int. Journal on Digital Libraries* (3) 2000: 115–130.

⁸¹ <https://huggingface.co/t5-base>

abstractively. Longer pieces of text must be split into smaller chunks, and then propagated to the model. Additionally, the model is about twice the size compared to models containing only an encoder (BERT) or decoder (GPT).

A model similar in some respects to vT5 is MBART⁸². It is a sequence-to-sequence auto-encoder model pretrained with BART objectives for many languages. VLMBART, a model designed for keyword generation, uses BART autoencoder architecture for keyword generation. Like the T5 model it is a generative model working both abstractively and extractively. It achieves similar results to the T5 model. In general, the comparison showed that bigger models achieve greater accuracy, hence the best results were for vT5-large, mBART-large, vT5-base, mBART-base (best → worst).

4.2.5 Keywords generation evaluation

Quantitative evaluation

The relevance and coverage of keyword prediction were compared using standard evaluation metrics like F1-score, precision, recall, and accuracy. We used both micro- and macro-precision and recall values, as well as their harmonic means F1-score averaged over the documents in the test set. We measured these scores at several ranks (k=1, 3, 5, and more) for each approach. We considered two different scenarios: a) using the full set of keywords assigned in the training and test set and b) training and/or evaluating only keywords that occur at least 10 times in the stratified dataset.

The highest F1 score of 0.335 (min. 10 words) and 0.227 (no limit) was achieved for vT5. The next place was taken by ExtremeText 0.145 and 0.094 respectively. Lastly, the TermoPL algorithm achieved F1 scores equal to 0.048 and 0.056, and in the last place KeyBert (with results below 0.01).

We additionally tested several other baseline keyword extraction approaches, including FirstPhrases and TopicRank⁸³, PositionRank⁸⁴, MultipartiteRank⁸⁵, TextRank⁸⁶, KPMiner⁸⁷ and Tfidf with some adjustments aimed at boosting their performance (such as lemmatizing input text). The results were less than 0.025 F1 on all ranks.

⁸² Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, “Multilingual Denoising Pre-training for Neural Machine Translation”, Transactions of the Association for Computational Linguistics 8, 2020: 726–742.

⁸³ A. Bougouin, F. Boudin, B. Daille, “TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction”, International Joint Conference on Natural Language Processing (IJCNLP), Nagoya 2013: 543-551, <https://hal.archives-ouvertes.fr/hal-00917969/>.

⁸⁴ C. Florescu, C. Caragea, PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents”, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver 2017: 1105-1115, <https://aclanthology.org/P17-1102/>.

⁸⁵ F. Boudin, “Unsupervised Keyphrase Extraction with Multipartite Graphs”, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans 2018: 667–672, <https://arxiv.org/abs/1803.08721>.

⁸⁶ R. Mihalcea, P., Tarau, “TextRank: Bringing Order into Texts”, <https://aclanthology.org/W04-3252.pdf>

⁸⁷ S. R. El-Beltagy, A. Rafea, “KP-Miner: Participation in SemEval-2”, Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala 2010: 190-193, <https://aclanthology.org/S10-1041.pdf>.

The details behind experiments are described in P. Pęzik, A.Mikołajczyk, A.Wawrzyński, et.al., *Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer*, *ACIIDS2022*.

Qualitative evaluation

In a qualitative evaluation of keyword generation tools, a team of 3 human annotators (bibliographers and literary scholars) assessed keywords and key phrases extracted from 1,000 digitised papers from a Polish scientific journal in literary studies (“Teksty Drugie”) from 1990–2000. The quality of the data varied. There were a lot of OCR errors in many texts. We did not perform additional data cleaning.

We tested two unsupervised approaches, an extractive one employing BERT embeddings and cosine similarity (KeyBERT) and an abstractive (or descriptive) one employing a ranking algorithm based on a knowledge base (Monte Carlo methods, Wikipedia2Vec). We also tested the vIT5 model on a smaller sample of texts. We generated summaries based on full texts.

The results from the extraction approach were the poorest. The vast majority of keywords were not accurate. The main issue was a strong dependence on document quality (i.e., OCR quality) leading, for example, to cutting off key phrases. With the descriptive approach we obtained far better representation of the content of the papers allowing their comparison. Some results were acceptable, others not. By far the best results were obtained with pIT5. The keywords were comprehensible and conveyed the main topic of the text well. The detailed results of the evaluation are presented in Table 4.2.1.

	Extractive approach	Descriptive approach	VIT5
Pros	inclusion of named entities	mostly comprehensible results	almost all keywords/keyphrases comprehensible, and almost no vocabulary outside the humanities/literature studies
	inclusion of idiosyncratic keywords, outside the controlled vocabulary	inclusion of literary terms automatic lemmatization inclusion of longer phrases summarizing the paper well enough (difficult to extract directly from the text)	inclusion of literary terms automatic lemmatization and capital letters capturing the most important topic inclusion of both keywords that generalize the content of the text, and keywords directly derived from the text
Cons/Issues	strong dependence on document quality (i.e., OCR quality), resulting, e.g., in cutting keywords off in the middle	overrepresentation of keywords/phrases outside of literary studies (e.g., words related to dance/music and religion, and disciplines such as astronomy or political science)	very general keywords mixed with very specific ones, usually accurate, but often overly detail-focused
	many keywords extracted from footnotes (Oxford citation style) or quotations, incl. literary quotations	keyword overlap, e.g., including hyponyms and hyperonyms	few words truncated (resulting from OCR errors), but a relatively small percentage
	many random multi-word expressions, sometimes cut off, sometimes too long	incorrect disambiguation: words/phrases constituting literary terms and also used outside the literary context are sometimes not recognised as literary terms too general keywords: often correctly identified keywords provide little about the specifics of an article, in turn, at a more detailed level we get less accurate results	significant share of buzzwords which can apply to a large pool of texts significant share of keywords strongly semantically related, sometimes synonyms, sometimes meronyms/hyponyms
		named entities very rarely recognized	some keywords occur in the text only once and are not relevant to the content insufficient coverage of named entities when using the Polish model (the model trained on English-language data returns more named entities) some words difficult to interpret and not very useful from an indexing point of view

Table 4.2.1. Overview of the results of the qualitative evaluation of three approaches to keyword extraction for Polish

4.2.6 Issues and challenges

Data quality

The key issue is the varying (and often low) quality of data, both the texts from which the keywords are to be extracted (i.e., scientific papers) and the training data. The keyword generation tool needs to work well for both born-digital and digitised texts, often containing OCR errors and requiring extensive preprocessing. There are still too few training datasets, and those available differ in terms of number of annotated keywords and annotation schemes adopted (sometimes described in detail, sometimes not).

Many texts, especially older ones, do not include abstracts. Whereas for more recent texts, we often only have an abstract. This is a significant difficulty, because sometimes keyphrases are not to be found explicitly in the text, or for instance they are available in the full-text but not in the abstract.

Multilingualism and multidisciplinary

Another problem is the multilingualism and multidisciplinary of datasets prepared for model training. The languages are sometimes incorrectly annotated or even mixed in both abstract and keyword fields. Thanks to the rapid development of deep learning language models, multilingualism is becoming much less of a problem, as multilingual models are getting more

commonly used. Moreover, new papers are more often published in English, rather than in local languages.

Multidisciplinarity results in a wide variety of abstract structure and keyword description approaches. In SSH, the structure of abstracts is very heterogeneous. Unlike, for example, in the medical sciences, there are usually no strict guidelines for creating abstracts in SSH scientific journals. As a result, each author may have his or her own strategy for creating an abstract and assigning keywords. The presence of specialised vocabulary also poses difficulties.

Evaluation

Evaluating models can also be a challenge. Exact matching of keyphrases is not a good approach, but there is no any better. Manual analysis is costly and time-consuming.

4.2.7 The needs of the scientific community

The results of the survey of Polish scientific publishers and editors of scientific journals on their use of digital solutions and the need for new services and tools indicate that – at least in the Polish scientific community – there is a relatively high demand for new metadata enrichment services and tools, including automatic keyword extraction. Publishers and editors of scientific journals are potential users of such solutions.

None of the respondents used automatic keywords generation tools. A large group was not aware of the existence of such solutions (Figure 4.2.1).

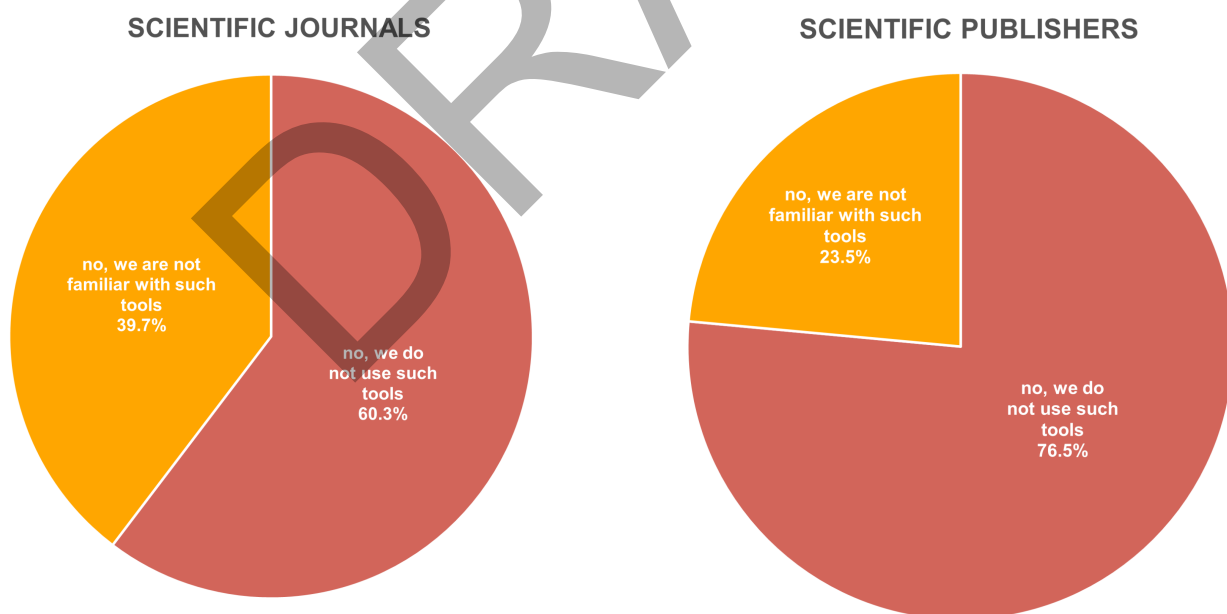


Figure 4.2.1. Use of automatic keyword extraction tools by editors of Polish scientific journals and publishers

More than half of the editors expressed interest in using new automatic keyword generation software, provided, however, that it was free (Figure 4.2.2). Interest in paid software was very low among this group.

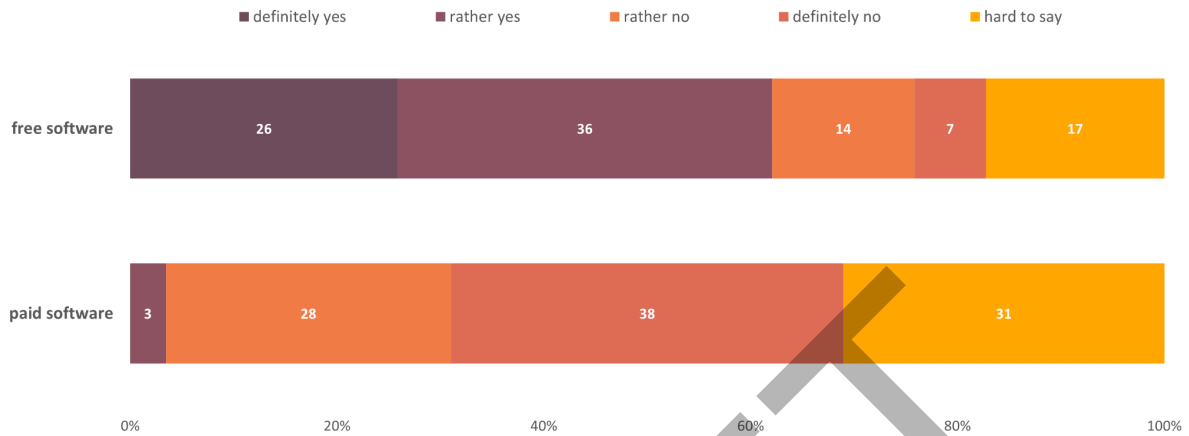


Figure 4.2.2. Interest among editors of Polish scientific journals in using new automatic keyword extraction software.

More interest was shown by scientific publishers (Figure 4.2.3). Almost 80% of respondents said they would use free metadata extraction software and 9% would be willing to pay for such a solution.

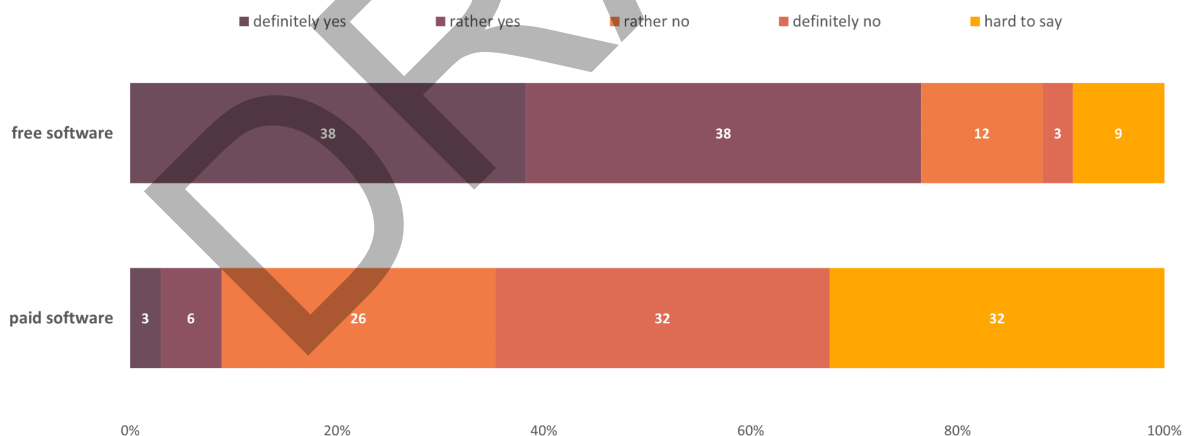


Figure 4.2.3. Interest among Polish scientific publishers in using new automatic keyword extraction software

Recommendations

Considering the needs of the scientific community and the evaluation of automatic keyword generation tools, we propose the following solutions:

- consider two levels of keywords: a more general one, allowing for the assignment of papers to disciplines, and a more specific one, within each discipline (e.g., sub-disciplines, styles, movements, themes)
 - use broad descriptors such as “philosophy”, “history”
 - extract the literary studies domain from wordnet⁸⁸ or deciding whether or not to include words/phrases based on their distance from the name of the discipline and the main subject of study, e.g. “literature”, “literary studies”
 - use domain-specific vocabularies and extending them via Wikipedia and wordnets
- add post-processing keyword aggregation methods for keyphrases that are too specific,
- add post-processing keyword ranking methods that will sort the keywords in the order from the most to the least important,
- include named entities, in particular, person names
- employ contrastive approaches to extract author-specific words/phrases absent from vocabularies (idiosyncratic terminology)
- improve the quality (esp. coherence) of the metadata of papers used to train models
- liaise with global open metadata and citation data initiatives such as I4OC and use their available sources to train models.

A key step to improve the performance of existing tools is to provide well-defined and carefully prepared training datasets: a corpora of text documents with manually assigned “perfect” keywords, after manual correction of OCR and other errors, with checked inter-annotator agreement scores.

⁸⁸ Ch. Fellbaum (ed.), *WordNet – An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.

4.3. Keywords in the TRIPLE Project

Cezary Rosiński⁸⁹

In addition to aggregating keywords originally assigned to articles by authors or providers, the project has also created a separate space for keyword extraction. The GoTriple Vocabulary⁹⁰ is a vocabulary of subjects in Social Sciences and Humanities (SSH) to be used by the GoTriple platform annotation service. For the annotation mechanism to be effective for publications of all the 9 languages that will be supported, the Vocabulary must contain a sufficient number of subject headings in form of semantic concepts and, at the same time, these concepts must have labels in as many of these languages as possible.

The project partners concluded that the most effective approach would be to establish a vocabulary by building upon an existing Linked Open Data vocabulary. The first step was to determine which vocabulary to use as the foundation. To accomplish this, the contributors of the task compiled a list of existing vocabularies⁹¹ and their main features, such as the number of concepts, the languages supported, the subject scope, and whether they sufficiently cover social sciences and humanities concepts, as well as whether they are published as LOD.

The GoTriple Vocabulary was initially developed by identifying 14 basic concepts from the Frascati taxonomy⁹² under SSH. Based on these, 37 broad concepts from the Library of Congress Subject Headings (LCSH) were identified. For each of these – with the usage of the Linked Data API of the Library of Congress⁹³ – their semantic Simple Knowledge Organization System (SKOS) representation was retrieved. For each of these the `skos:narrower` property was followed and their children in SKOS were extracted with 2513 concepts in total.

The multilingualism of the Vocabulary was increased by 1) following links of LCSH 2) following links to Wikidata and extracting labels in various languages and 3) ingesting existing mappings from national vocabularies (French and Italian National Libraries) to LCSH. The multilingualism was further enhanced by using an automatic translation service to produce missing labels which then were validated and/or curated by partners. As a result, the coverage increased significantly (e.g. for Greek from 9.67% to 88.73% and Polish from 13.10% to 99.84%).

⁸⁹ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-6136-7186>.

⁹⁰ <http://semantics.gr/authorities/vocabularies/SSH-LCSH>

⁹¹ H. Georgiadis, M. Błaszczynska, M. Maryl. TRIPLE Deliverable: D2.4 Report on identification and creation of new vocabularies. Zenodo. 2023, <https://doi.org/10.5281/zenodo.7539922>.

⁹² <https://www.oecd.org/innovation/frascati-manual-2015-9789264239012-en.htm>

⁹³ APIs for LoC.gov available at <https://loc.gov/apis>.

		October 2020		June 2021	
		# of Concepts with labels in language (out of 2565)	language coverage percentage	# of Concepts with labels in language (out of 2565)	language coverage percentage
Greek	el	248	9.67%	2276	88.73%
French	fr	1624	63.31%	2219	86.51%
Polish	pl	336	13.10%	2561	99.84%
German	de	1028	40.08%	2561	99.84%
Italian	it	686	26.74%	2560	99.81%
Portuguese	pt	347	13.53%	2560	99.81%
Spanish	es	406	15.83%	2560	99.81%
Croatian	hr	153	5.96%	2228	86.86%

Table 4.3.1. Progress in enhancing multilingualism in GoTriple Vocabulary

The GoTriple Vocabulary needed to meet certain key requirements in order to be published as a LOD vocabulary in SKOS, which is the standard data model for concept-based vocabularies. This includes being accessible via persistent URIs in both HTML and RDF formats, following the SKOS data model, and being published under an open licence for anyone to use. Additionally, the vocabulary was required to be continuously updated by the Triple consortium, including adding new labels, links to other vocabularies, and concepts. To ensure these requirements were met, it was decided to host the GoTriple Vocabulary on the dedicated platform Semantics.gr.

Semantics.gr⁹⁴ is based on a state-of-the-art infrastructure that supports the development, curation and interlinking of vocabularies, thesauri, classifications, classification schemes and authority files – altogether called Vocabularies – and their publication as LOD. The infrastructure is being developed in-house by the National Documentation Centre in Greece (EKT, a Greece TRIPLE partner) and has been used for cataloguing and enrichment data. It employs semantic knowledge representation technologies.

The novelty of Semantics.gr lies in the fact that, besides SKOS, it can support any Data Model that can be expressed as an OWL ontology. The Data Model chosen in each case generates the template based on which a Vocabulary Schema is created. The Vocabulary Schema, in turn, stipulates in detail the structure of the vocabulary that will be created in the infrastructure. In practice, the Vocabulary Schema defines in detail the entry/update form through which the user adds or updates Semantic Resources.

⁹⁴ <https://www.semantics.gr/authorities/info/semanticsPage?language=en>



Figure 4.3.2. The GoTriple Vocabulary in Semantics.gr

Semantics.gr aspires to serve as a central public platform for publishing trustworthy LOD vocabularies, especially scientific terminology and authority files, that can be further used by any third party in order to enhance the quality and interoperability of their digital resources.

The main goal of the GoTriple platform as a discovery system is to aggregate all possible SSH data and provide sufficient query functionalities. Therefore, keywords in GoTriple primarily function as supporting mechanisms for cataloguing information. The service created a rich controlled vocabulary for SSH stored in SKOS format that uses LOD structures. It is possible to enrich the controlled vocabulary both within headings, using links to further services, and within the entire vocabulary, which can be enriched with relevant concepts. However, the GoTriple platform also presents original author-generated keywords alongside each text. Since these keywords exist as strings, they cannot be used in search mechanisms. Moreover, they are not semantically linked to the controlled vocabulary. In the future, it would be worthwhile to develop a mechanism for using original keywords and enriching the controlled vocabulary with them. The use of artificial keywords, on the other hand, can help to describe aggregated material more comprehensively, and mechanisms known, for example, from Annif software can offer a combination of the effects of automatic keyword generation with the existing semantic environment of the GoTriple website.

4.4. Guidelines for keywords as humanities research data in the context of TRIPLE – summary

- 1) The humanities data providers are more susceptible to content description – via keywords – that is not controlled in terms of applied vocabularies (as smaller data providers are playing a role in the humanities and the content is more language-specific). In this case, **a dedicated effort to map existing keywords onto relevant controlled vocabularies** is needed to allow for more interoperability between resources. This effort should take into account **multiple languages, especially smaller ones**, that are not present in existing, popular controlled vocabularies.
- 2) Controlled vocabularies for the humanities need to be multilingual.
- 3) At least some resources are being **described by keywords of various origins**, in multiple places (in the humanities this is even more systemic because many monographs' content is being independently indexed through librarian services). In order to create a research dataset, a level of preprocessing of these resources is needed in the form of gathering different keywords (while attributing the method of their creation), analysing and comparing them. This is a prerequisite for relevant mapping and enrichment of those keywords onto existing vocabularies.
- 4) Due to the growing number of publications that demand content description, there is **a need to develop automated solutions for keywords description**. This automation is especially needed in the humanities where historical content is relatively more relevant. Humanities deal with digitised historic content that cannot be thematically described by humans on a large scale.

5. Abstracts

5.1. Abstracts in the SSH

Bianca Kramer⁹⁵

5.1.1 Role of abstracts

Abstracts of research publications (including but not limited to journal articles, book chapters, books and monographs) fulfil an important role in scholarly communication: they help readers determine the relevance of the publication, communicate key findings and summarise the content of the publication. They also increase visibility of the work through their inclusion in bibliographic databases, which helps discovery. Finally, they are useful in classifying and grouping texts, for instance in creating subject classifications.

Importantly, these benefits still hold when the publication itself is open access, as easy access to abstracts facilitates discovery and selection. In addition, both for search and classification, using full text is not always preferable to using abstracts, both for practical and conceptual reasons.

While books and monographs might often not include a formal abstract as is customary for many journal articles and book chapters, they often have a short description available to help readers quickly appraise the topic and content. For the topic of this chapter, these are considered equivalent to abstracts.

5.1.2 Importance of open abstracts

Abstracts can be considered part of the metadata of publications, and as such, are an important part of open science. For understanding and making informed decisions based on research in a transparent way, it is not only important for data and publications themselves to be openly available, but also for the metadata of these outputs to be openly available and reusable, as well as having open analytical tools and applications to use.

As GoTriple aims to provide infrastructure for social sciences and humanities to make this possible, it benefits from the availability of open metadata, including abstracts.

5.1.3 Ways abstracts are (made) available

Abstracts are usually available from both publisher websites and bibliographic databases. However, there are several limitations to the availability of abstracts in these ways.

Publishers usually make abstracts free to read on their websites, but while these abstracts can be read by humans who arrive on the individual publication's landing page, the large-scale processing of the same content by machines is difficult. For search and retrieval

⁹⁵ Sesame Open Science, Open Abstracts, <https://orcid.org/0000-0002-5965-6560>.

purposes, abstracts included in bibliographic databases are more useful. These databases, however, often have restrictions on access (for instance, only accessible with an institutional licence) and reuse restrictions. In addition, many have a limited disciplinary scope.

5.1.4 Initiative for Open Abstracts (I4OA)

In 2020, the Initiative for Open Abstracts (I4OA)⁹⁶ was started as a collaboration between scholarly publishers, infrastructure organisations, librarians, researchers and other interested parties to advocate and promote the unrestricted availability of the abstracts of the world's scholarly publications in trusted repositories where they are open and machine-accessible.

While I4OA is not prescriptive in how and where to open up abstracts, it does ask publishers, where possible, to (also) submit them to Crossref.

5.1.5 Crossref as centralised infrastructure for metadata

Crossref provides a centralised infrastructure for making abstracts openly available (together with other publication metadata) in a machine-readable way with minimal restrictions. While obviously limited to publications with a Crossref DOI, for these publications Crossref is a centralised source of uniformly formatted, authorised metadata. Having abstracts available as part of these metadata opens them up to reuse by many applications and research infrastructures.

Such downstream usage can be direct, e.g. knowledge extraction for subject classification, screening abstracts for inclusion in systematic reviews, and enriching metadata in institutional repositories (IRs) and research information systems (CRIS). Usage can also be indirect, e.g. through inclusion in bibliographic databases, use in bibliographic mapping tools (to create network visualisations around research topics) and applications that employ machine learning based on abstracts (like Scholarcy⁹⁷ and ASReview⁹⁸)⁹⁹.

Publishers that are members of Crossref can submit abstracts for journal articles, books and book chapters, conference papers, posted content, dissertations, reports, and standards. Abstracts can be deposited through various ways: as part of the XML submitted to Crossref, by using a web deposit form, or a platform-specific plugin, like the one developed for OJS. Like other Crossref metadata, they are then available through the public Crossref APIs for download and reuse.

⁹⁶ <https://i4oa.org/>

⁹⁷ <https://www.scholarcy.com/>

⁹⁸ <https://asreview.nl/>

⁹⁹ See A. Tay, B. Kramer, L. Waltman, Why openly available abstracts are important — overview of the current state of affairs, <https://medium.com/a-academic-librarians-thoughts-on-open-access/why-openly-available-abstracts-a-re-important-overview-of-the-current-state-of-affairs-bb7bde1ed751>.

What about copyright?

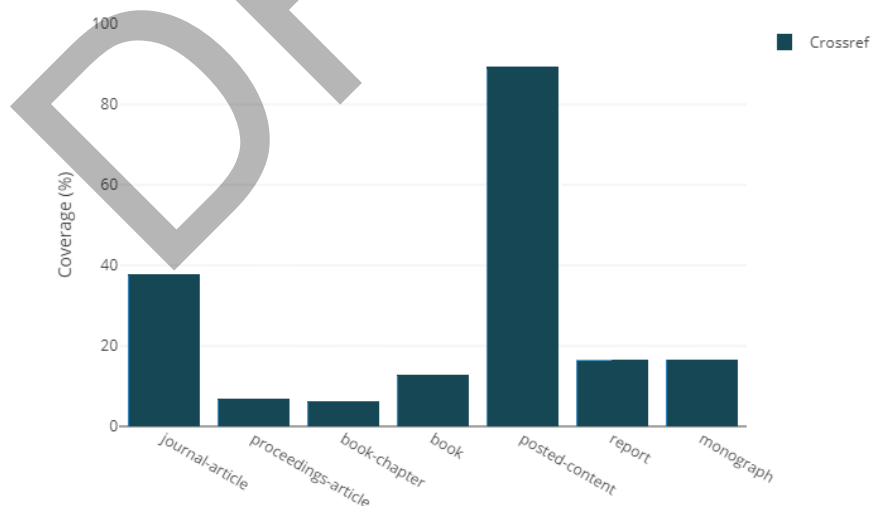
Abstracts have a somewhat unique position: while functionally, they can be considered part of a publication's metadata, they are also in themselves creative textual output and as such, can be subject to copyright by either the authors, their institution or the publisher.

In practice, (at least in the European Union) abstracts can freely be used for text- and data mining, under the TDM exception for academic use in the European Copyright Directive. However, they cannot be republished without permission, unless covered by the licence the publication itself is shared under, e.g. Creative Commons licences used by many open access publications. It is important to note that the above is meant to provide some guidance and clarification regarding reuse of abstracts, but does not constitute legal advice.

The restrictions mentioned here also apply to abstracts made available through Crossref as part of a publication's metadata. Crossref itself states that it “generally provides metadata without restriction; however, some abstracts contained in the metadata may be subject to copyright by publishers or authors” (see e.g. ¹⁰⁰)

Current availability of open abstracts in SSH in Crossref

Coverage of abstracts in Crossref greatly differs between publication types. As shown in Figure 5.1.1, abstracts are available for close to 40% of recent journal articles (published between 2020 and 2022), and preprints have an extremely high coverage of 80%. For recent books and monographs though, abstract coverage is lower than 20%, and for book chapters even lower than 10%.



*Figure 5.1.1. Abstract coverage in Crossref per publication type (publication years 2020-2022).*¹⁰¹

¹⁰⁰ <https://www.crossref.org/documentation/retrieve-metadata>

¹⁰¹ Data source here and for the figures below: Crossref metadata (author's own analysis).

Absence of abstracts in Crossref can either be a function of publishers not providing them as part of publication metadata (either for all their content, or for a subset of journals), or the publication not having an abstract in the first place. For journal articles, we can see a clear distinction between providers who choose to submit abstracts to Crossref (and often formally support I4OA) and those that don't (see Figure 5.1.2). For SSH, the latter notably include CAIRN and Project Muse. However, even publishers that do provide abstracts to Crossref, often don't reach 100% abstract coverage. Especially for journal publishers in social sciences and humanities (like Brill and Erudit in Figure 5.1.2), this could be affected by the proportion of journal publications that do not contain abstracts, including, for instance, book reviews (in addition to editorials, letters to the editor and other non-journal content that can be found in journals in all disciplines).

DRAFT

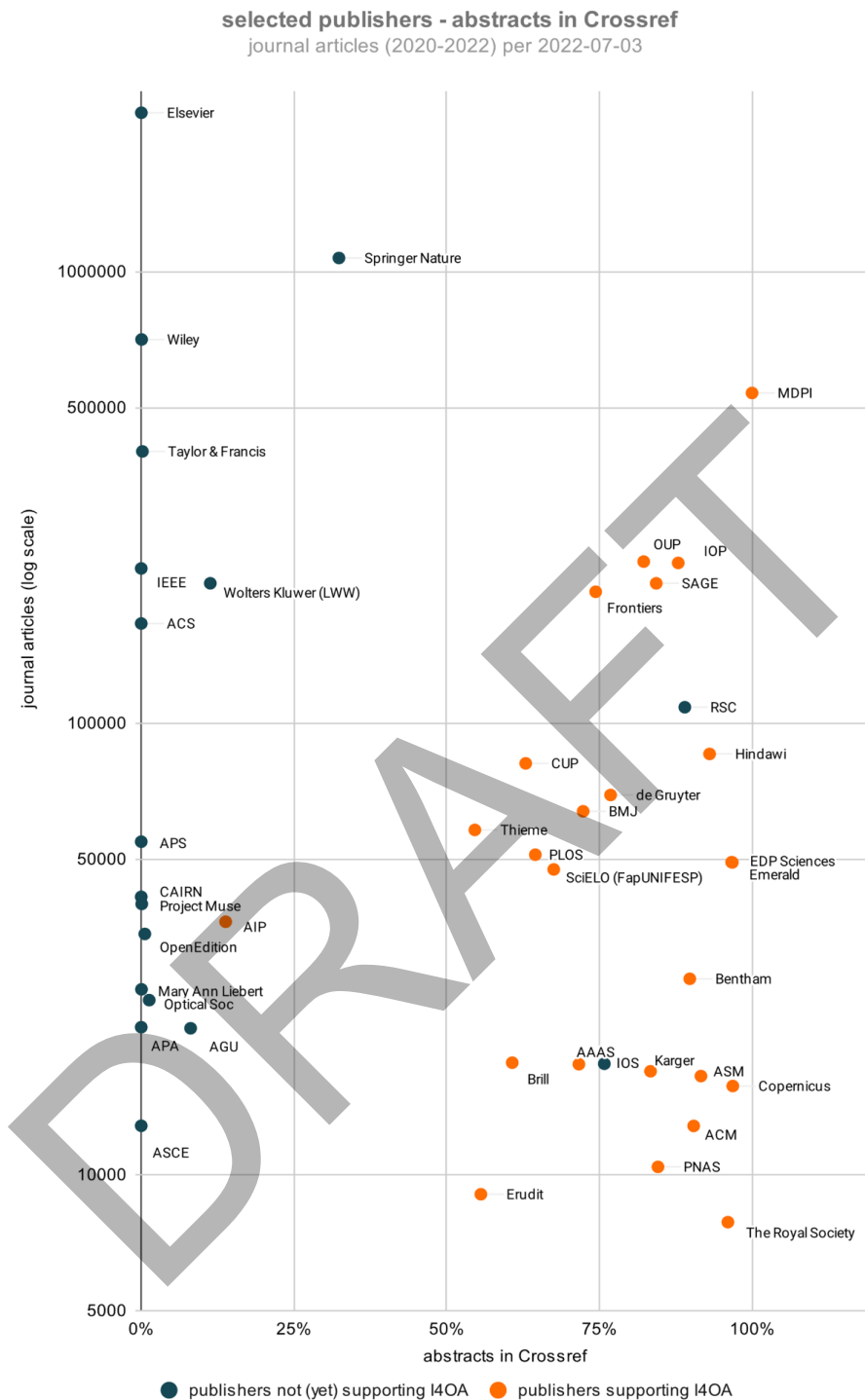


Figure 5.1.2. Abstract coverage in Crossref for journal articles – selected publishers (publication years 2020-2022)

For books, book chapters and monographs, the picture that emerges is even more binary. Most larger publishers, even those that support I4OA and supply abstracts for articles, do not supply abstracts for long-form publications. Only very few (e.g. Oxford University Press (OUP), IGI Global) do it for almost all their long-form content (Figure 5.1.3). In addition to the availability of abstracts for these publication types and/or technical barriers in publisher

workflows, there might be limited awareness among book publishers of the possibility and importance of supplying abstracts as part of Crossref metadata.



Figure 5.1.3. Abstract coverage in Crossref for books, book chapters and monographs – selected publisher (publication years 2020-2022)

Even though books, book chapters and monographs so far have been taken together, there are interesting differences in abstract coverage for these publication types between publishers, which cannot be explained by overall larger availability of abstracts for one of these publication types over the other. For instance, while Oxford University Press (OUP) supplies abstracts for close to 100% of both their books and book chapters, Cambridge University Press only supplies abstracts for books and monographs, not book chapters (Figure 5.1.4). Incidentally, these differences in coverage for a given publisher also influences their overall percentage of abstract coverage for long-form materials as depicted in Figure 5.1.3. For instance, because Cambridge University Press publishes many book chapters, for which they do not supply abstracts, their overall abstract coverage for long-form materials is only 6%, masking the fact that they do, in fact, provide abstracts for the majority of their books and monographs.

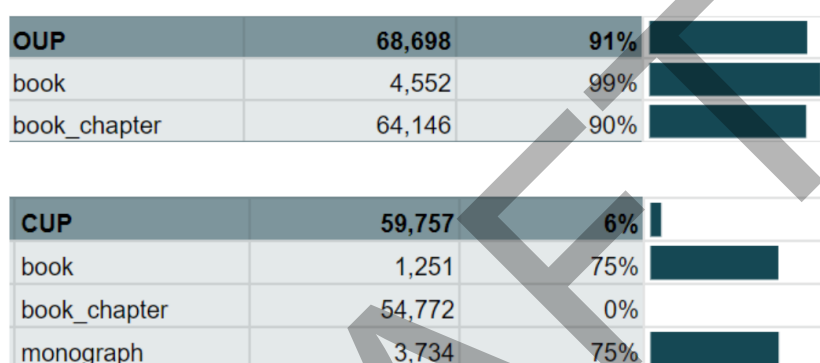


Figure 5.1.4. Abstract coverage in Crossref for books, book chapters and monographs separately - Oxford University Press (OUP) and Cambridge University Press (CUP) (publication years 2020-2022)

It is important to not only look at the larger (book) publishers, but also take into account the situation for mid-size and smaller publishers. One reason is the diversity in abstract coverage for long-form publications among smaller publishers. Two examples for Polish institutions are given in Figure 5.1.5.

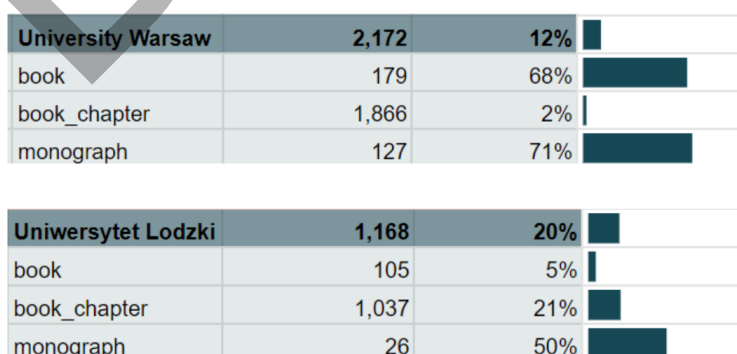


Figure 5.1.5. Abstract coverage in Crossref for books, book chapters and monographs separately - University of Warsaw and Uniwersytet Lodzki (publication years 2020-2022)

There are also some excellent examples of good practice among mid-size and smaller publishers, for instance Berghahn Books (who also formally support I4OA), Amsterdam University Press (AUP) and Bologna University Press, which all provide abstracts for the large majority of their long-form publications, irrespective of publication type (Figure 5.1.6).

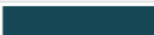
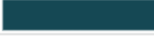








Berghahn Books	89	100%	
book	52	100%	
monograph	37	100%	
AUP	2,267	92%	
book	286	100%	
book_chapter	1,949	92%	
monograph	32	69%	
Bologna UP	109	88%	
book	103	88%	
monograph	6	83%	

Figure 5.1.6. Abstract coverage in Crossref for books, book chapters and monographs separately - Berghahn Books, Amsterdam University Press (AUP) and Bologna University Press (publication years 2020-2022)

The final reason to pay attention to abstract coverage for smaller publishers (and include them in awareness-raising and outreach actions) is the potential knock-on effect on availability of abstracts in downstream applications. If attention is solely focused on increasing abstract availability for larger publishers, then those are the abstracts that will be available in downstream services as well, potentially leading to a serious bias in visibility and usage of abstracts.

For any Crossref member, a visual representation of metadata coverage (including abstract coverage) can be found in Crossref Participation Reports¹⁰² for various publication types including journal articles and book chapters. Unfortunately, for books, abstracts are currently not included in the metadata types shown.

5.1.6 Some issues around (open) abstracts in SSH

Persistent identifiers

The previous paragraphs discussed a workflow through which abstracts for a.o. articles, books, book chapters and monographs can be made available for reuse via Crossref. As mentioned, this is an approach suitable for publications that have a Crossref DOI, which is still less common for long-form publications than for journal articles. Even though the value of persistent identifiers is well-established (both for identification and linking purposes, as well as for use of the associated metadata), DOIs, and among them Crossref DOIs, are not the only persistent identifiers available, and a sole focus on Crossref DOIs therefore

¹⁰² <https://www.crossref.org/members/prop/>

excludes many publications. Especially for smaller and non-Western publishers, cost and required technical expertise are sometimes felt to be prohibitive for participation in Crossref (notwithstanding Crossref's efforts to remedy this through its sponsorship program and community engagement activities).

Another source of publication metadata (including abstracts) is through harvesting information through OAI-PMH endpoints – both from publishers that enable this, and from institutional repositories. For instance, OpenEdition makes all its abstracts available through OAI-PMH, even if they do not supply them through Crossref. Many aggregators make use of OAI-PMH to harvest metadata, including OpenAIRE, BASE (Bielefeld Academic Search Engine), CORE and GoTriple itself. It would be interesting to compare coverage and quality of metadata (including abstracts) retrieved via OAI-PMH and via Crossref, both across the landscape of SSH output as well as for specific publishers that supply metadata via both routes. In practice, infrastructures like GoTriple could consider enriching the metadata harvested through OAI-PMH with metadata from DOI registrars like Crossref for those publications that have DOIs.

Presence of abstracts

As mentioned earlier, books and monographs might often not include a formal abstract as is customary for journal articles and book chapters. However, they often do have a short description available to help readers quickly appraise the topic and content, which thus function as an abstract. It will depend on the publisher as well as the metadata schema used whether this text is included as abstract in the metadata. For instance, Crossref has a dedicated metadata field for abstracts, while the Dublin Core standard (often used for OAI-PMH and also mapped to the metadata scheme for DataCite DOIs) contains a more general metadata field labelled 'description' which can be used for abstracts, but also for other information (either with or without labels using controlled vocabulary).

The above illustrates the confluence of factors that affect the availability and potential for retrieval of abstracts in metadata for long-form publications – running the gamut from disciplinary norms to technical implementation.

Language diversity

To properly account for language diversity, abstracts for publication output should be captured in all languages in which they are made available. This is important for all areas of research, not just SSH, although language diversity is arguably most prominent in SSH. Availability of abstracts in multiple languages depends both on inclusion criteria of research infrastructures as well as on publishers' practices in including abstracts in metadata. In this respect, outreach and awareness efforts should purposely include non-English providers.

In the case of multilingual abstracts for one publication, it is important that all language versions are included in the publication's metadata. While metadata schemes allow for this, in practice this is not always done correctly. One example discussed during the booksprint for this report concerned metadata for the same publication, provided by the publisher in

Dublin Core¹⁰³ and JATS XML¹⁰⁴ format with both an English and Polish abstract, while the Crossref metadata¹⁰⁵ for the same publication contained only the English abstract. This is not due to limitations in Crossref metadata, since Crossref encourages members to include titles and abstracts in multiple languages in your metadata¹⁰⁶, with it being confirmed via personal communication that this is true for all content types that accept abstracts). Here again, awareness and where needed, capacity support for providers seems crucial to ensure optimal availability of metadata.

Other sources

As discussed, publication metadata, including abstracts, can be sourced via multiple ways, including metadata associated with DOIs and metadata harvested through OAI-PMH. Additionally, aggregators might make agreements to receive metadata from publishers directly and/or employ web scraping to collect and/or enrich metadata. These methods will vary in transparency and provenance as well as coverage and completeness. In addition, the extent to which aggregators make their metadata, including abstracts, available for reuse might also differ in accessibility (e.g. presence of paywalls) and interoperability. Therefore, it is useful to assess and compare metadata coverage and completion for various aggregators, both as an end-user as well as an aggregator or provider (like GoTriple) looking for sources to enrich existing metadata. For metadata of research output in SSH, some potential sources include WorldCat, BASE, OpenAIRE, DataCite (which includes metadata from Zenodo) and OpenAlex.

As an example of the potential added value of multiple metadata sources, Figure 7 shows the coverage of abstracts for various publication types in OpenAlex, both for Crossref DOIs (Figure 7A) and for research output without DOI in OpenAlex (Figure 7B). As shown, OpenAlex does have abstracts for a considerable proportion of journal articles and book chapters, and to a somewhat lesser extent also for books and monographs, for which Crossref does not have abstracts. In addition, OpenAlex can to some extent also be a source of abstracts for long-form publications without DOIs.

¹⁰³See: https://bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:bibliotekanauki.pl:1968869.

¹⁰⁴See: <https://bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=jats&identifier=oai:bibliotekanauki.pl:1968869>.

¹⁰⁵ See: <http://api.crossref.org/works/10.17651/SOCJOLING.35.2>.

¹⁰⁶ <https://www.crossref.org/documentation/principles-practices/best-practices/multi-language/>

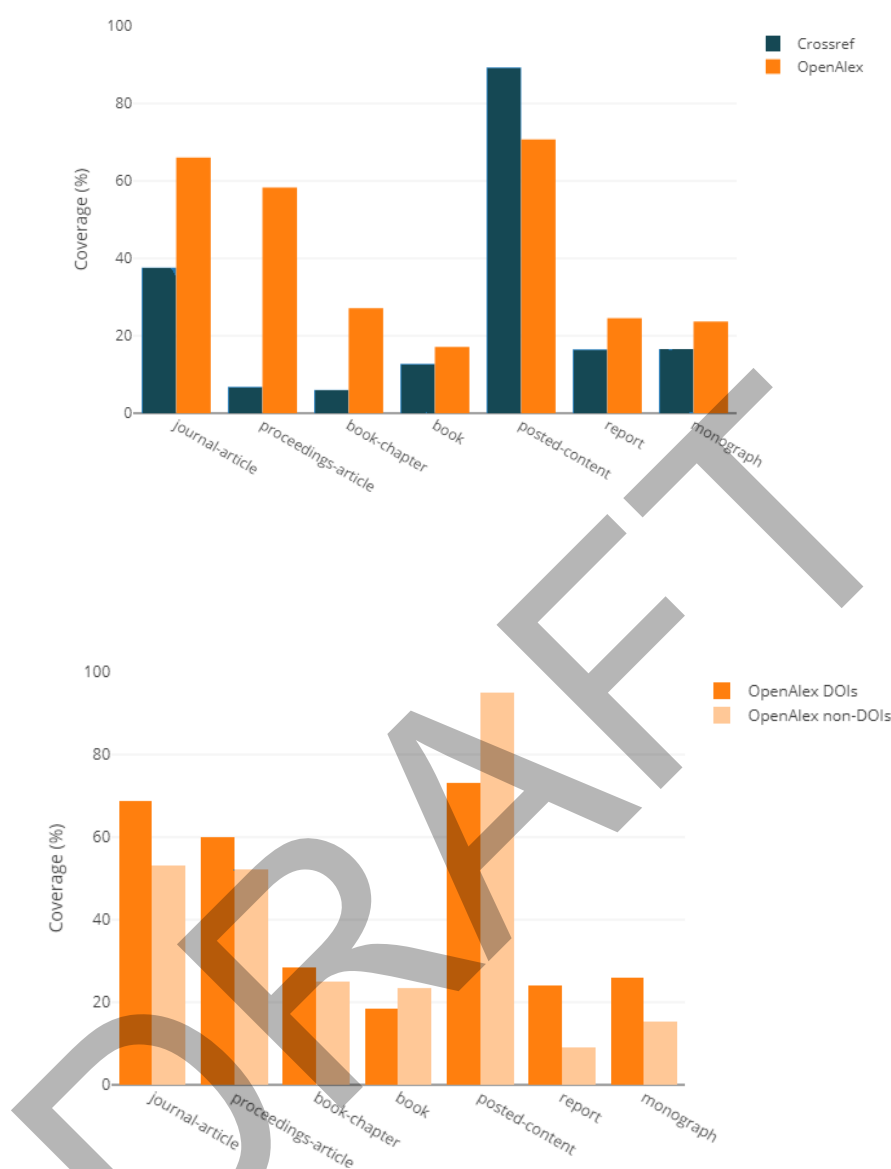


Figure 5.1.7A, 5.1.7B). Abstract coverage for Crossref DOIs in Crossref and OpenAlex (A) and for publications with or without DOI in OpenAlex (B) per publication type (publication years 2020-2022, sampled June 2022).

5.1.7 Conclusions and Implications for GoTriple

Open availability of abstracts and other metadata is, in principle, independent of platform/infrastructure. Ideally, publishers and other providers would make abstracts available as part of metadata for all their publications in machine-readable format in all their endpoints, to allow maximal availability and re-use.

In practice, both disciplinary norms (predominant publication types and presence of abstracts for these publication types) and social, financial and technical barriers limit the availability of abstracts as metadata.

For aggregators like GoTriple, coverage of abstracts will therefore depend on how and where metadata are harvested (eg. directly via OAI-PMH endpoints or via Crossref). A combined approach, where metadata are enriched by harvesting from additional sources, might provide added value.

To prevent, as much as possible, bias in the metadata that are collected and made available, careful attention should be paid to inclusion of the long tail of publishers as well as to accommodation of language diversity, including multilingual abstracts.

DRAFT

5.2. Highway to abstract. The present and the future of automatically generated abstracts

Agnieszka Karlińska¹⁰⁷, Cezary Rosiński¹⁰⁸, Nikodem Wołczuk¹⁰⁹

5.2.1 Missing abstracts

As stated in the previous subchapter (*Abstracts in the SSH*), one of the challenges in the current SSH data ecosystem is the lack of abstracts in the form of a structured metadata field. This can happen when the abstract has not been created for a particular document at all or when the information is present in the document (such as a PDF file), but has not been inserted into the metadata schema for its description. This is also the case for the GoTriple providers which will be investigated closely below and followed by a discussion of possible technological challenges and solutions that could be applied to improve the abstract coverage for the platform.

In the case of one of GoTriple's providers – Library of Science (*Bibliotekanauki.pl*) – this happens in 20% of cases. So, around 90 000 documents (out of more than 490 000) are missing an abstract in their metadata description.

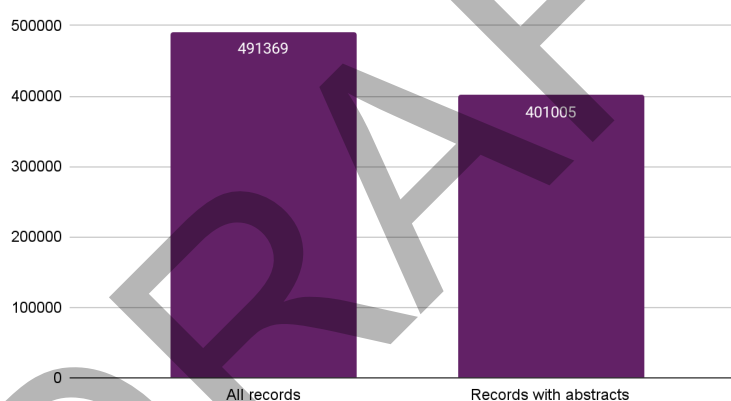


Figure 5.2.1. Articles with abstracts in Library of Science (*bibliotekanauki.pl*, September 2022).

The case of the Library of Science, however, is an exception. The high number of released abstracts is due to the service's focus on aggregating current scientific materials which are usually better equipped with metadata than older resources. Services that collect materials from a wider time range do not have such a rich representation of abstracts in their data. Old materials requiring digitalization often suffer from a lack of such metadata. A case of such a service is the Digital Repository of Scientific Institutes (*RCIN*¹¹⁰), which among 105 287 records with the type "text" has only 8984 indexed abstracts. That's a mere 8.5%.

¹⁰⁷ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-4846-7086>.

¹⁰⁸ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-6136-7186>.

¹⁰⁹ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-4303-2016>.

¹¹⁰ <https://rcin.org.pl/>

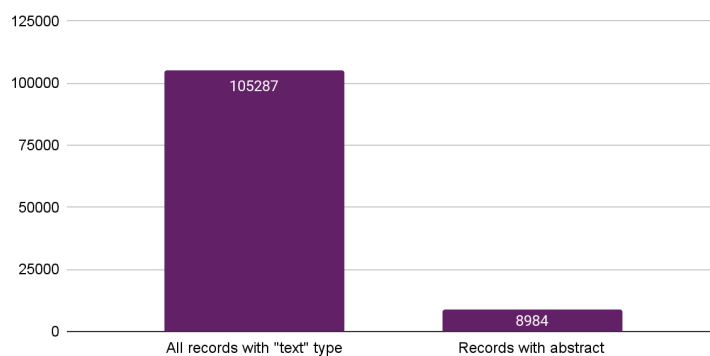


Figure 5.2.2. Records with abstracts in Digital Repository of Scientific Institutes RCIN (rcin.org.pl, September 2022).

The GoTriple dataset demands inclusion of not only contemporary scientific output, but also older materials originating from the GLAM sector where abstract coverage is limited.

5.2.3 Making up for missing abstracts

There are two complementary approaches to supplementing missing abstract data: metadata extraction and metadata generation¹¹¹ (Figure 5.2.3).

The first strategy applied on a digitised item without abstract metadata involves Optical Character Recognition, Optical Layout Recognition and Semantic Segmentation procedures with creating a textual layer of the digitised document, identifying abstract and extracting its content, followed by the addition of the extracted abstract as a new metadata type.

The second strategy indicates a different set of difficulties. Where an abstract is not available, and that kind of situation may be common for older documents, there is a need to generate an abstract from scratch. The traditional activity which involves authorial factor or even adding that new piece of metadata by professional bibliographers or librarians is rather unlikely to happen because of financial constraints. Therefore the only possible solution is to use NLP and Machine Learning to do so.

¹¹¹ Here we leave out the possibilities of manual creation of abstracts that are missing due to how time- and cost-consuming this solution would be.

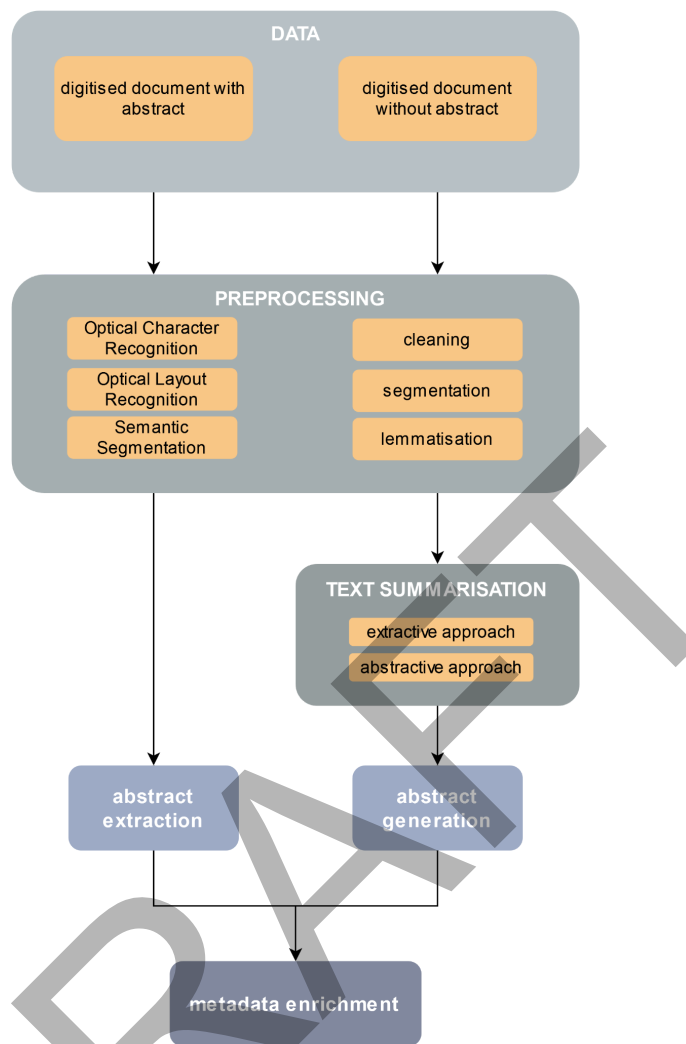


Figure 5.2.3. Supplementing missing abstract workflow

In NLP, the task of creating a shorter text that is the best possible semantic representation of the original document – i.e. including the most important or relevant information – is referred to as ‘automatic text summarization’. Research on text summarization has been carried out since the 1950s. The goal is to develop a system for machine-generated summaries that are equivalent to human-generated ones¹¹².

Similar to keyword generation systems, text summarization systems are classified into two types: extractive and abstractive¹¹³. In an extractive approach, paragraphs, sentences or phrases are obtained from the source document and then put together to produce the summary¹¹⁴. In abstractive techniques text is paraphrased and the summary contains sentences and phrases different from the source document. Abstractive methods are usually highly complex as they require full-text comprehension and extensive natural language

¹¹² M. Gambhir, V. Gupta. “Recent automatic text summarization techniques: a survey”, *Artif Intell Rev*, 47, (2017): 1–66.

¹¹³ G. Sharma, D. Sharma, “Automatic Text Summarization Methods”, *A Comprehensive Review. SN COMPUT. SCI*, 4 (2022), 33.

¹¹⁴ A. Khan, N. Salim, “A review on abstractive summarization methods”, *Journal of Theoretical and Applied Information Technology*, 59.1 (2014): 64-72.

processing¹¹⁵. Therefore, the research community has long focused primarily on improving extractive techniques¹¹⁶.

Extractive text summarization has become a popular field within NLP, resulting in a whole range of approaches implementing several machine learning and optimization techniques. They perform different types of clustering that usually aim to extract the most diverse topics occurring in the original document¹¹⁷. For this reason, extractive summarization is said to rely solely on sentence scoring to maximise the topical coverage and minimise redundancy, while coherence (i.e. the extent to which ideas are related to each other) and cohesion (i.e. textual fluency) are to be considered only in the case of abstractive summaries¹¹⁸. It is not always the case as there are extractive methods allowing for both sentence scoring and text fluency¹¹⁹.

Since extractive text summarization is a relatively well-developed field, current research is increasingly shifting toward abstract and hybrid text summarization methods¹²⁰. One factor is recent advances in neural methods, which provide a feasible framework for obtaining an abstract representation of the meaning of the original text¹²¹. At the core of abstractive summarization, there are three pipelined tasks: information extraction which gets useful information from text using noun or verb phrases, content selection which works by selecting a subset of important phrases from the extracted text, and surface realisation, which combines selected words or phrases in a sequence by using grammatical rules and lexicons¹²². Abstractive text summarization methods are divided into structure-based approaches, semantic-based approaches, and deep learning-based approaches¹²³ or linguistic approaches, e.g. information-based or tree-based methods, and semantic approaches, e.g. template-based methods or ontology-based methods¹²⁴. Neural networks-based text summarization is considered state-of-the-art¹²⁵.

All summarization methods and models, both extractive and abstractive, share the common goal of generating summaries that are informative, non-redundant, and coherent. Despite several improvements, they still pose some issues and challenges. One of the key challenges for abstractive text summarization is ensuring factual consistency, i.e. including in the summary only those statements that can be derived directly from the original

¹¹⁵ A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, I. M. De Rosal Setiadi. 2020. "Review of automatic text summarization techniques & methods", *Journal of King Saud University – Computer and Information Sciences*, 34 (4) (2020): 1029-1046.

¹¹⁶ M. Gambhir, V. Gupta. "Recent automatic text..."

¹¹⁷ I. Okulska, "Team Up! Cohesive Text Summarization Scoring Sentence Coalitions", L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J. M. Zurada (eds), *Artificial Intelligence and Soft Computing. ICAISC 2020. Lecture Notes in Computer Science*, vol 12416. Springer, Cham 2020.

¹¹⁸ C. C. Aggarwal, "Machine Learning for Text". Springer, Cham 2018.

¹¹⁹ I. Okulska, "Team Up! Cohesive Text Summarization Scoring Sentence Coalitions", L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J. M. Zurada (eds), *Artificial Intelligence and Soft Computing. ICAISC 2020. Lecture Notes in Computer Science*, vol 12416. Springer, Cham 2020.

¹²⁰ A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, I. M. De Rosal Setiadi. 2020. "Review of automatic text..."

¹²¹ H.-C.g Lin, V. Ng, "Abstractive Summarization", *A Survey of the State of the Art. Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01) (2019): 9815-9822.

¹²² H.-C.g Lin, V. Ng, "Abstractive Summarization"...

¹²³ S. Gupta, S. K Gupta, "Abstractive summarization", *An overview of the state of the art, Expert Systems with Applications*, 121 (2019): 49-65.

¹²⁴ A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, I. M. De Rosal Setiadi. 2020. "Review of automatic text..."

¹²⁵ S. Gupta, S. K Gupta, "Abstractive summarization"...

document¹²⁶. Recent studies show that about 30% of abstractive summaries generated by neural network sequence-to-sequence models involve fact fabrication such as containing names that never appeared in the original text¹²⁷. The main challenge from the perspective of generating abstracts of scientific texts is the strong dependence on the domain (in this case the scientific discipline), which makes it impossible to develop a fully universal approach. Multilingual text summarisation is also a challenge. New developments are emerging in this field, but there are still a limited number of datasets for low/mid-resource languages¹²⁸.

5.2.4 Who needs this data and what for

The GoTriple dataset needs the fullest possible abstract coverage due to both facilitating the needs of the end user of its public interface and the value they bring for understanding society, culture and science. The reasons for having abstracts were already discussed in detail in previous sections. Abstracts are critical for the enrichment workflow GoTriple employs, especially keyword attribution, generation of data visualisations such as Streamgraphs and Knowledge Maps.

From this point of view it is critical to tackle abstract extraction and generation, especially when we take into account the fact that many types of documents – such as books, book chapters, non-article documents within journals – lack them, which was discussed by Bianca Kramer in the section entitled *Abstracts in the SSH*. The question remains – who should tackle this challenge and on which stage of document lifecycle? GoTriple is the high-level aggregator of contents created by its providers and other aggregators. The need for a larger abstract coverage is also well-understood down the pipeline – by scientific journals and publishers.

In a survey¹²⁹ conducted within the Dariah.lab¹³⁰ project respondents (Polish scientific journals and publishers) answered questions about their knowledge of automatic metadata extraction tools and their willingness to use them. Although 70% of publishers and 60% of scientific editors responded that they add or enrich metadata of their publications, none of the publishers and only two scientific journals used automatic techniques. What is more, a large group of respondents was not aware of the existence of automatic metadata extraction tools. The vast majority of editors expressed interest in using new automatic metadata extraction software, provided, however, that it was free. Interest in paid software was very low among this group (Figure 5.2.4).

¹²⁶ F. Nan, R. Nallapati, Z. Wang, C. Nogueira dos Santos, H. Zhu, D. Zhang, K. McKeown, B. Xiang, „Entity-level Factual Consistency of Abstractive Text Summarization”, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume 2021: 2727–2733, Online. Association for Computational Linguistics.

¹²⁷ W. Kryscinski, B. McCann, C. Xiong, R. Socher, „Evaluating the Factual Consistency of Abstractive Text Summarization”, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 9332–9346.

¹²⁸ T. Hasan, A. Bhattacharjee, M. Saiful Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. Sohel Rahman, R. Shahriyar. 2021. „XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages”, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021: 4693–4703, Online. Association for Computational Linguistics.

¹²⁹ <https://operas.pl/2022/04/22/pytamy-redakcje-czasopism-i-wydawnictw-naukowych-o-technologie/>

¹³⁰ <https://lab.dariah.pl/en/>

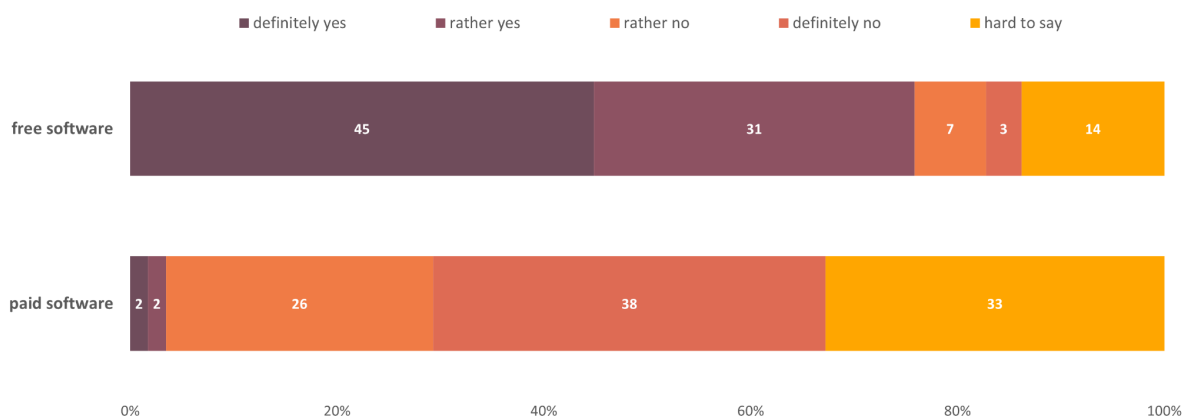


Figure 5.2.4. Interest among editors of Polish scientific journals in using new automatic metadata extraction software

More interest was shown by scientific publishers (Figure 5.2.5). Over 80% of respondents said they would use free metadata extraction software and 24% would be willing to pay for such a solution.

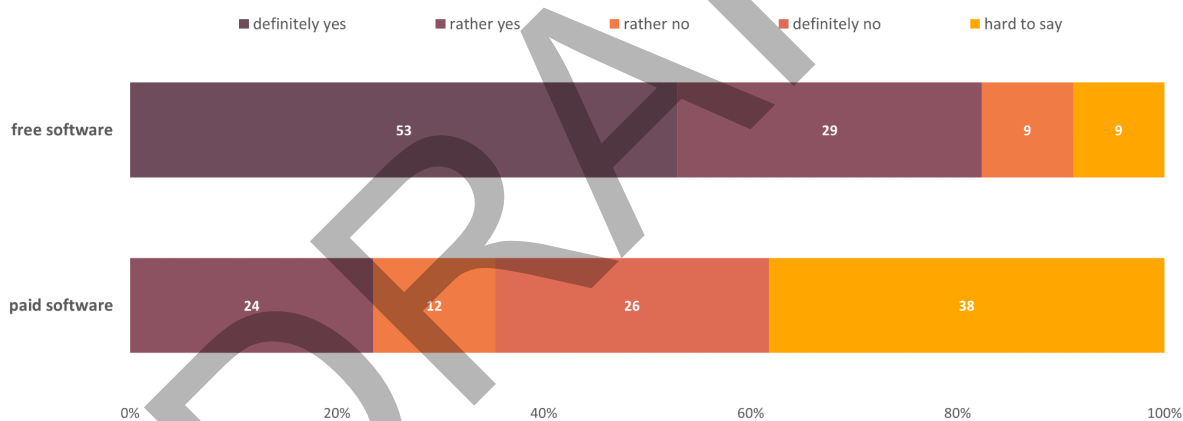


Figure 5.2.5. Interest among Polish scientific publishers in using new automatic metadata extraction software

None of the respondents used abstract generation tools. As in the case of metadata extraction, a large group was not aware of the existence of such solutions.

For scientific journal editors, there was slightly less interest in using a new service to automatically generate abstracts than there was for metadata extraction tools, but more than half of respondents were still willing to use such a service, provided it was free (Figure 5.2.6).

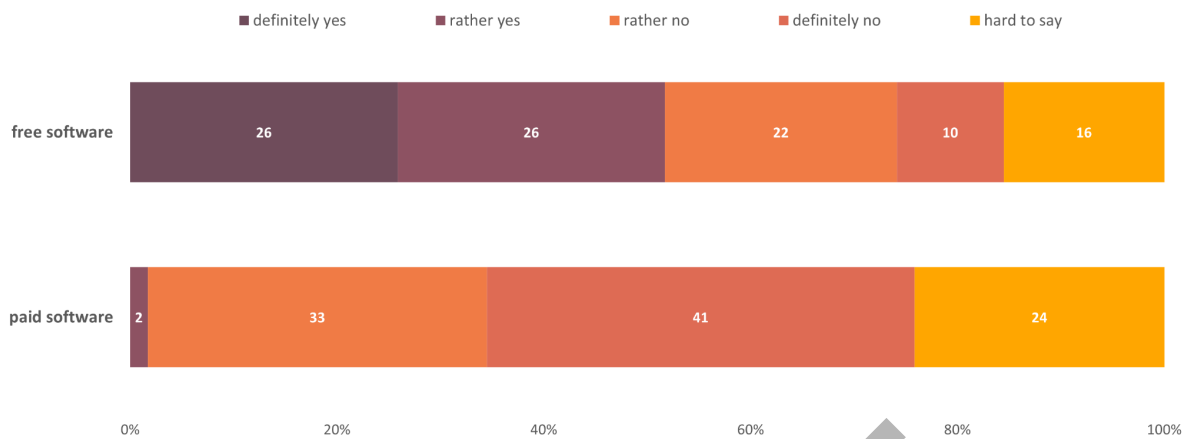


Figure 5.2.6. Interest among editors of Polish scientific journals in using new abstract generation tools

Again, scientific publishers showed a bit more interest. 70% of respondents would use a free service and 12% a paid one (Figure 5.2.7).

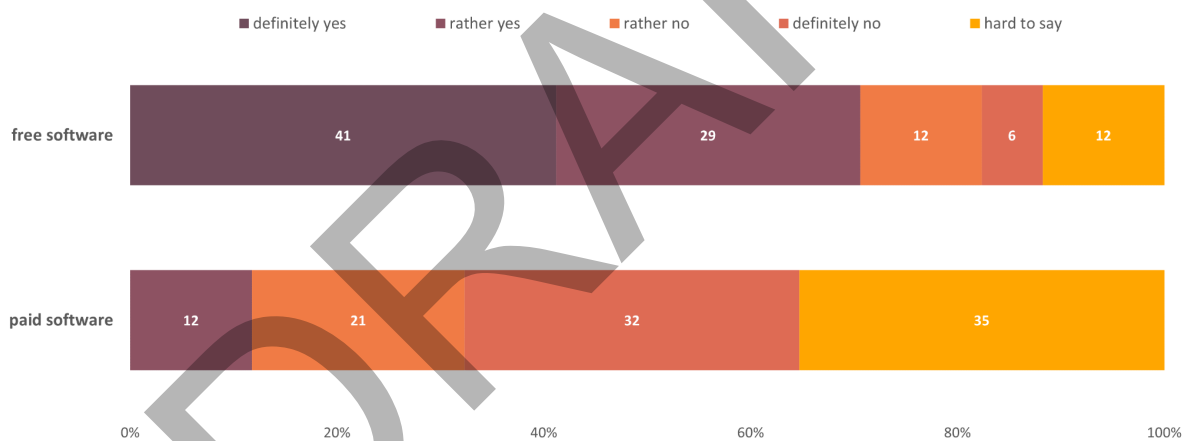


Figure 5.2.7. Interest among Polish scientific publishers in using new abstract generation tools

5.3. Guidelines for abstracts as humanities research data in the context of TRIPLE – summary

- 1) Humanities need to **take full advantage of existing infrastructure for open metadata** – such as Open Abstracts – which includes for example registering abstracts through Crossref services. Through this **humanities data becomes more accessible and easier to link to different data spaces**.
- 2) As different engines and algorithms rely on abstracts – for example for automated keywords extraction – **there is a need to supplement abstracts if they do not exist**. This also might apply to non-scientific documents, such as cultural ones, important for the humanities. Hence the relevance of automated approaches to text summarisation and the possibilities of applying those methods in creation of large research datasets.
- 3) With the development of NLP methods and resources, such as large corpora, **text summarisation can provide insight into documents that cannot be accessed in their entirety** (provided the copyright considerations and regulations regarding data mining allow for it).
- 4) Multilingual aspects of abstracts are especially important for the humanities – this means both capturing all language versions of abstracts, but also leveraging the potential for automated translations of abstracts into all necessary languages (and current developments in the NP and AI might be of great help here).

6. Citations

6.1. Transparency meets open citations

Silvio Peroni¹³¹

In the scholarly ecosystem, a bibliographic citation is a *conceptual* directional link from a citing entity to a cited entity, used to acknowledge or ascribe credit for the contribution made by the author(s) of the cited entity. Citations are one of the core elements of scholarly communication. They enable the integration of our independent research endeavours into a global graph of relationships that can be used, for instance, to analyse how scholarly knowledge develops over time, assess scholars' influence, and make wise decisions about research investment.

However, as citation data, i.e. pieces of *factual* information aiming at identifying entities and relationships among them, are of great value to the scholarly community, it has been a “scandal”¹³² that they have not been recognised as part of the commons. Indeed, only recently we have seen some efforts – such as the – already discussed in the previous chapter – Initiative for Open Citations (I4OC¹³³) – that have tried to change the *behind-the-paywall status quo* enforced by the companies controlling the major citation indexes used worldwide, convincing scholarly publishers to support the unrestricted availability of scholarly citation data by publishing them in suitable open infrastructures, such as Crossref¹³⁴ and DataCite¹³⁵.

Of course, as for many other kinds of data, putting bibliographic and citation data behind a paywall is a threat to enabling the full reproducibility of research studies based on them (e.g. in bibliometrics, scientometrics, and science of science domains), even when such studies are published in open access articles. For instance, the results of a recent open access article by Spinaci *et al.* (2022)¹³⁶ published on *Digital Scholarship in the Humanities*, which aimed to analyse the citation behaviour of Digital Humanities (DH) research across different proprietary and open citation databases, are not fully reproducible since the majority of the databases used – namely Scopus, Web of Science, and Dimensions – do not make their bibliographic and citation data openly available.

In addition, the coverage of publications and related citations in specific disciplines, in particular those within the Social Sciences and the Humanities (SSH), is not adequate

¹³¹ University of Bologna, OpenCitations, <https://orcid.org/0000-0003-0530-4305>.

¹³² D. Shotton, “Open citations” *Nature*, 502 (7471) (2013): 295–297. <http://dx.doi.org/10.1038/502295a>.

¹³³ <https://i4oc.org>

¹³⁴ <https://crossref.org>

¹³⁵ <https://datacite.org>

¹³⁶ G. Spinaci, G. Colavizza, S. Peroni, “A map of Digital Humanities research across bibliographic data sources”, *Digital Scholarship in the Humanities*, 2022, fqac016. <https://doi.org/10.1093/lc/fqac016>.

compared to that of other fields¹³⁷. Usually, this is due to the limited availability of born-digital publications accompanied by a wide variety of publication languages, publication types (e.g. monographs), and complex referencing practices that may limit their automatic processing and citation extraction. As a side effect, such a partial coverage may result in a considerable bias when analysing SSH disciplines compared to STEM disciplines that usually have better coverage in existing citation databases.

All these scenarios have at least another negative effect on the area strictly concerned with the research assessment, which often uses quantitative metrics based on citation data to evaluate articles, people, and institutions. Indeed, the unavailability and partial coverage of bibliographic and citation data create an *artificial* barrier to the transparency of the processes used to decide the careers of scholars in terms of research, funding, and promotions.

In the past years, several initiatives around the world have highlighted the importance of reforming research assessment exercises, such as those summarised in Figure 6.1.1: the French National Plan for Open Science¹³⁸, the San Francisco Declaration on Research Assessment¹³⁹, the Leiden Manifesto for Research Metrics¹⁴⁰, and the recent proposal for a reform of the research assessment system by the European Commission (2021)¹⁴¹ that is becoming formalised under the Coalition for Advancing Research Assessment (COARA)¹⁴². All these initiatives agree on a few essential characteristics necessary for having a trustful assessment system:

- to be open and transparent by providing *machine-readable, unrestricted* and *reusable* data and methods for calculating the metrics used in research assessment exercises, and
- to leave to the *research community*, instead of commercial players, the control and ownership of the crucial infrastructures and tools used to retrieve, use and analyse such data within research assessment systems.

Thus, the leading guideline that can be abstracted is to follow Open Science practices *even when assessing research* and not only when performing research.

¹³⁷ G. Colavizza, S. Peroni, M. Romanello, “The case for the Humanities Citation Index (HuCI)”, A citation index by the humanities, for the humanities. *International Journal on Digital Libraries*, (2022), <https://doi.org/10.1007/s00799-022-00327-0>; A. Martín-Martín, M. Thelwall, E. Orduna-Malea, E. Delgado López-Cózar, “Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI”, A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), (2020):, 871–906. <https://doi.org/10.1007/s11192-020-03690-4>; V. K. Singh, P. Singh, M. Karmakar, J. Leta, P. Mayr, “The journal coverage of Web of Science, Scopus and Dimensions”, A comparative analysis. *Scientometrics*, 126(6) (2021): 5113–5142, <https://doi.org/10.1007/s11192-021-03948-5>; M. Visser, N. J. van Eck, L. Waltman, L. “Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic”, *Quantitative Science Studies*, 2(1) (2021): 20–41, https://doi.org/10.1162/qss_a_00112.

¹³⁸ <https://www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018/>

¹³⁹ <https://sfdora.org>

¹⁴⁰ <http://www.leidenmanifesto.org/>

¹⁴¹ European Commission, “Towards a reform of the research assessment system”. Scoping report. (KI-09-21-484-EN-N) 2021, Publications Office. <https://doi.org/10.2777/707440>.

¹⁴² <https://coara.eu/>

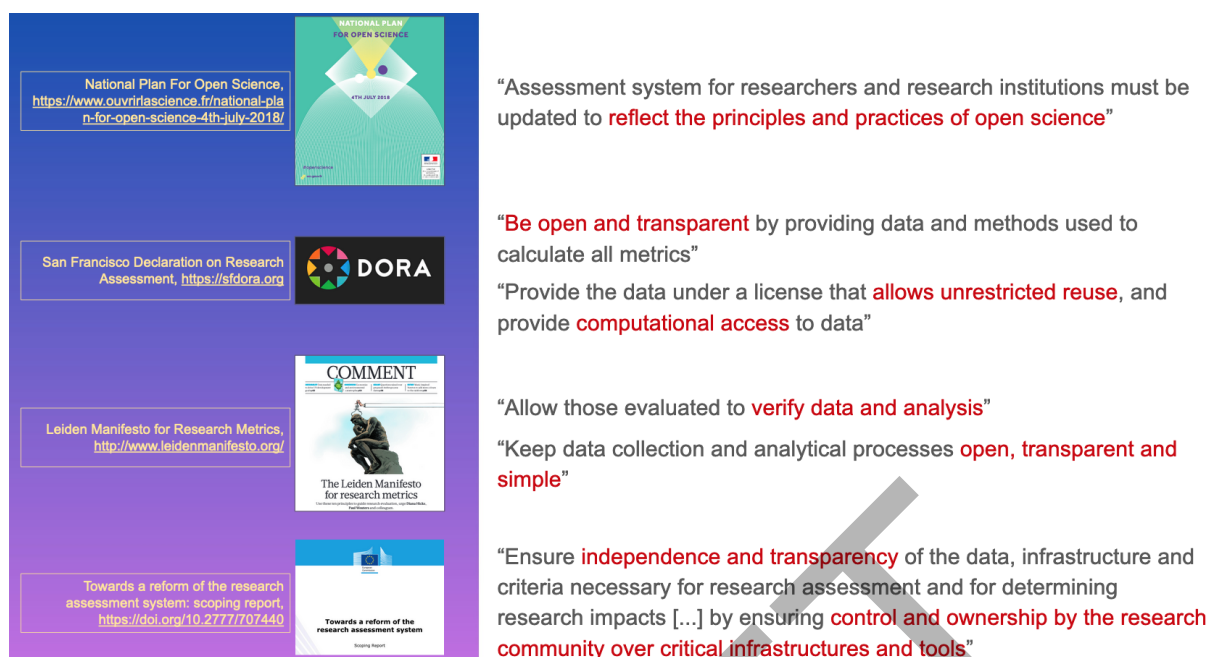


Figure 6.1.1. Some initiatives pushing for reforming the principles behind research assessment systems.

Within this context, OpenCitations¹⁴³ plays an important role, acting as a key infrastructure component for global Open Science, and pushes for actively involving universities, scholarly libraries and publishers, infrastructures, governments and international organisations, research funders, developers, academic policy-makers, independent scholars and ordinary citizens. The mission of OpenCitations is to harvest and openly publish accurate and comprehensive metadata describing the world's academic publications and the scholarly citations that link them, with the greatest possible global coverage and subject scope, encompassing both traditional and non-traditional publications, and with a breadth and depth that surpasses existing sources of such metadata, while maintaining the highest standards of accuracy and accompanying all its records with rich provenance information, and providing this information, both in human-readable form and in interoperable machine-readable Linked Open Data formats, under open licences at zero cost and without restriction for third-party analysis and re-use.

For OpenCitations, *open* is the crucial value and the final purpose. It is the distinctive mark and founding principle that everything OpenCitations provide – data, services and software – is open and free and will always remain so. OpenCitations fully espouses the aims and vision of the UNESCO Recommendations on Open Science, complies with the FAIR data principles, and promotes and practises the Initiative for Open Citations recommendation that citation data, in particular, should be Structured, Separable, and Open.

The most important collection of such open citation data is COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations¹⁴⁴. The last release, dated August 2022, contains more than 1.36 billion citation links between more than 75 million bibliographic entities that can be

¹⁴³ <https://opencitations.net>

¹⁴⁴ <https://w3id.org/oc/index/coci>

accessed programmatically using its REST API, queried via the related SPARQL endpoint, and downloaded in full as dumps in different formats (CSV, JSON, and RDF).

At the end of 2019, **OpenCitations** was selected by the **Global Sustainability Coalition for Open Science Services (SCOSS)** for their second round of crowd-funding support “aligns well with open science goals, is an innovative service, and if successful could be a game changer by challenging established proprietary citation services”

SCOSS stated that OpenCitations



Figure 6.1.2. Collaborations between OpenCitations and other Open Science infrastructures and services.

In addition to the publication of citation data, a considerable effort has been dedicated to collaborating with other Open Science infrastructures working in the scholarly ecosystem, as summarised in Figure 6.1.2. Since 2020, OpenCitations has significantly benefited from the scholarly community that resulted from the 2019 selection by the Global Sustainability Coalition for Open Science Services (SCOSS¹⁴⁵) of OpenCitations as a scholarly infrastructure worthy of financial support. The community funding permitted the appointments of people dedicated to the administration, communication, community development, and maintenance and improvement of the OpenCitations software and the computational infrastructure on which it runs. In addition, OpenCitations started its involvement with OpenAIRE¹⁴⁶ and the European Open Science Cloud (EOSC¹⁴⁷), and it is collaborating with other funded projects project such as RISIS2¹⁴⁸, OutCite¹⁴⁹, OPTIMETA¹⁵⁰ and BISON¹⁵¹.

While OpenCitations is currently providing a good set of citation data, which is already approaching parity with other commercial citation databases¹⁵² and that has already been used in a few studies for research purposes, there is still margin for improvement. Currently, the citations included in the OpenCitations Indexes come mainly from Crossref data, one of

¹⁴⁵ <https://scoss.org>

¹⁴⁶ <https://www.openaire.eu/>

¹⁴⁷ <https://eosc-portal.eu/>

¹⁴⁸ <https://www.risis2.eu/>

¹⁴⁹ <https://excite.informatik.uni-stuttgart.de/>

¹⁵⁰ <https://projects.tib.eu/optimeta/en>

¹⁵¹ <https://projects.tib.eu/bison/en/project/>

¹⁵² A. Martín-Martín, Coverage of open citation data approaches parity with Web of Science and Scopus, <https://opencitations.hypotheses.org/1420>.

the biggest open reference providers. However, Crossref does not cover all the publishers of DOI-based resources. Indeed, other DOI providers, in some cases, expose citation relations in their metadata, such as DataCite¹⁵³. In addition, DOI-based publications represent just a limited set of all the bibliographic entities published in the scholarly ecosystem. Other identifier schemas have been used to identify bibliographic entities – and, for some publications, there do not exist identifiers at all!

Thus, to address these two issues, OpenCitations is working on expanding its coverage in two different directions. On the one hand, OpenCitations is developing two new citation indexes of open references based on the holdings of DataCite and the National Institute of Health Open Citation Collection¹⁵⁴, which, together with COCI, will be cross-searchable through the Unifying OpenCitations REST API¹⁵⁵.

On the other hand, OpenCitations has started working to create a new database entitled OpenCitations Meta, which will provide three major benefits. First, it will permit storing in-house bibliographic metadata for the citing and cited entities involved in all OpenCitations Indexes, including author identifiers using ORCID and VIAF identifier schemes where available. Second, it will provide better query performance than the present API system, which obtains bibliographic metadata on-the-fly by live API calls to external services, such as Crossref and DataCite APIs. Finally, it will permit indexing citations involving entities lacking DOIs, by providing them OpenCitations Meta Identifiers.

This last collection, combined with automatic tools for citation extraction from digital formats, is crucial for increasing the coverage of underrepresented disciplines and fields in bibliographic databases, such as SSH publications. One of the OpenCitations' goals is to reduce this gap in citation coverage by setting up crowdsourcing workflows for ingesting missing citation data from the scholarly community (e.g. libraries and publishers). In the future, another contribution will be to set up tools for automatic extraction of citations that can also support small and local publishers, crucial assets for SSH research, that may find difficulties in carrying out citation extraction tasks on their own since using and maintaining a tool (or paying a company addressing those tasks on behalf of the publisher) requires extra costs beyond publishers' finances.

To conclude: OpenCitations is one piece of a puzzle that is working to change existing scholarly practices to create an open and inclusive future for science and research in which the scholarly community owns and is responsible for its own data.

¹⁵³ <https://datacite.org>

¹⁵⁴ B. I. Hutchins, K. L. Baker, M. T. Davis, M. A. Diwersy, E. Haque, R. M. Harriman, T. A. Hoppe, S. A. Leicht, P. Meyer, G. M. Santangelo, "The NIH Open Citation Collection", A public access, broad coverage resource. PLOS Biology, 17(10) (2019), e3000385. <https://doi.org/10.1371/journal.pbio.3000385>.

¹⁵⁵ <https://w3id.org/oc/index/api/v1/>

6.2. Citation data and GoTriple's data providers

Cezary Rosiński¹⁵⁶, Tomasz Umerle¹⁵⁷, Nikodem Wołczuk¹⁵⁸

6.2.1 Missing or lost citations?

Citations presence is sensitive to the format of data offered by the provider. In the case of one of GoTriple's providers – Library of Science¹⁵⁹ – Dublin Core data does not expose citation data. They are present, on the other hand, in the JATS format, in almost 60% of the articles:

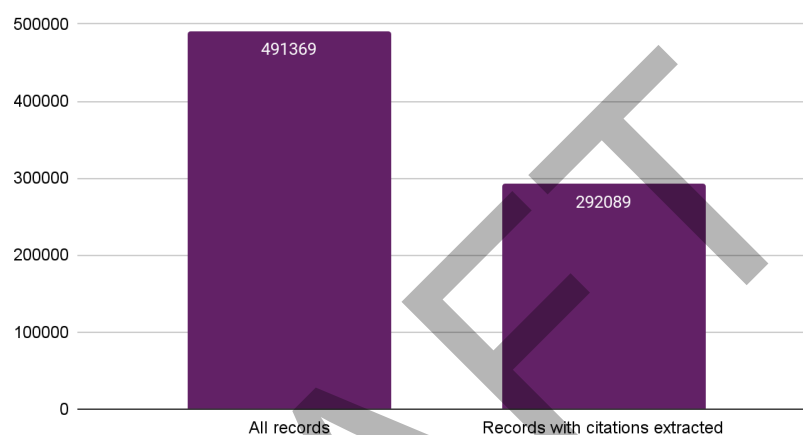


Figure 6.2.1. Citations in Library of Science in JATS format (*bibliotekanauki.pl*, September 2022).

Citation presence in the local service does not guarantee, however, that this information is carried to other services, such as Crossref. The analysis of the Library of Science's data prepared for the purposes of this report confirms that even though a certain metadata entity (record of the document) exists in Crossref, not every piece of original information is preserved there. For instance, metadata in the Library of Science in JATS format¹⁶⁰ may be equipped with extracted citations, the record presented in Crossref¹⁶¹ does not contain such information. What is worth mentioning, the exact same record in the Library of Science database in JATS format has citations extracted, but does not contain that part of metadata in DC format¹⁶², even though it is technically possible. In the end we are dealing with the open access aggregator of scientific output that already exposes citation data, but only locally.

¹⁵⁶ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-6136-7186>.

¹⁵⁷ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-7335-0568>.

¹⁵⁸ Institute of Literary Research, Polish Academy of Sciences (IBL PAN), <https://orcid.org/0000-0002-4303-2016>.

¹⁵⁹ bibliotekanauki.pl

¹⁶⁰ See: <https://bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=jats&identifier=oai:ibibliotekanauki.pl:1968869>.

¹⁶¹ See: <https://api.crossref.org/v1/works/10.17651%2FSOC.JOLING.35.2>.

¹⁶² See: https://bibliotekanauki.pl/api/oai/articles?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:ibibliotekanauki.pl:1968869.

From the perspective of the GoTriple dataset it means that – even if GoTriple aims to aggregate and expose citation data – 1. reliance on Dublin Core data is not a viable solution, 2. analysing different local formats offered by multiple providers to identify which one of them stores citation data will be a complex task, 3. relying on dedicated services that aggregate citation data (such as Crossref) does not guarantee even the minimum of exposing the same amount of metadata that is present locally.

6.2.2 Missing citations: extraction and enrichment

When the structured citation data is truly missing in the whole citation data lifecycle we are left with two solutions depending on the document and metadata quality.

The first case is a document without any citation metadata when we need to extract it from the body of text. Second case is when the document’s metadata contains a bibliography or reference list provided either as plain text or a set of divided citations, we need to parse and enrich this data.

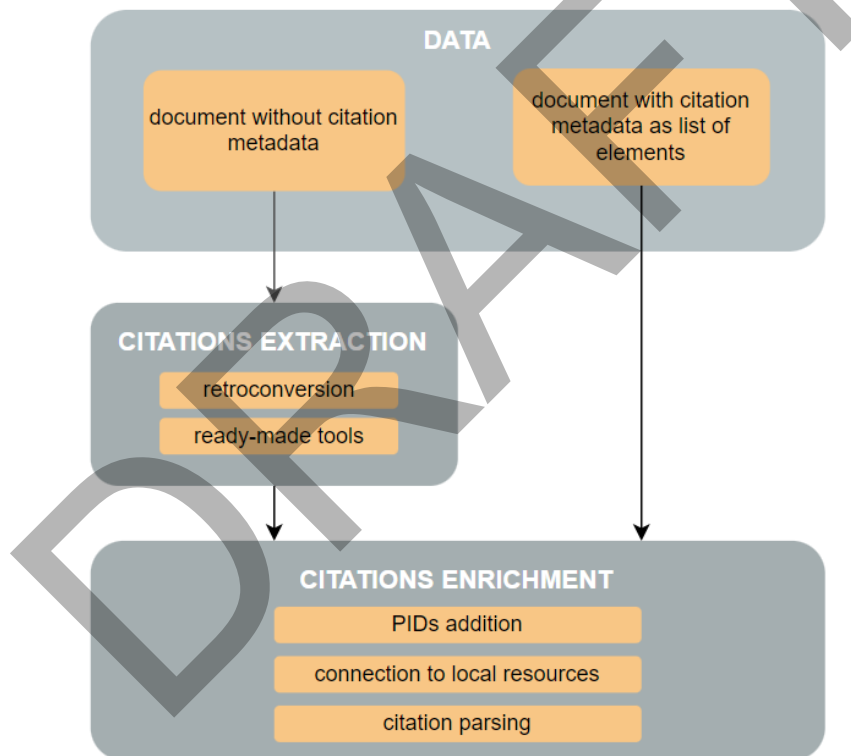


Figure 6.2.2. Missing citations - possible solutions.

Citations extraction relies on such software using OCR, OLR and Semantic Segmentation to identify parts of the document as references and enhance metadata with such information.

Citations enrichment is applied to a segmented set of extracted references. The further structuring of each reference may be provided by three strategies.

The first one consists of adding Persistent Identifiers to bibliographic information, such as DOIs. For instance, this is being done with the Crossref plug-ins for Open Journal System¹⁶³.

¹⁶³ See: <https://www.crossref.org/documentation/content-registration/ojs-plugin/>, <https://docs.pkp.sfu.ca/crossref-ojs-manual/en/references>

This solution is limited as it depends on the scope of Crossref, Worldcat and similar resources, but it remains especially useful for contemporary research output from journals as DOI attribution is mostly focused on them

The second solution requires connection with external resources such as national catalogues, bibliographies or scientific repositories. Their scope might be more suited for local needs – and they are not limited to the records with full texts. These authority services could provide knowledge at the local level via URIs attached to the references. That kind of linking may better shape the regional landscape of scientific research and prepare the information for the international exchange. But this solution would require efforts from local communities to provide enrichment tools.

The third way to enrich citations requires automatic parsing and may be provided, for instance, with the OJS plugins such as ParsCit Citation Extractor or ParaCite Citation Extractor. However, the most promising one is AnyStyle.io¹⁶⁴ – the free parser of references regardless of citation style which structures bibliographic data thanks to machine learning heuristics based on Conditional Random Fields.

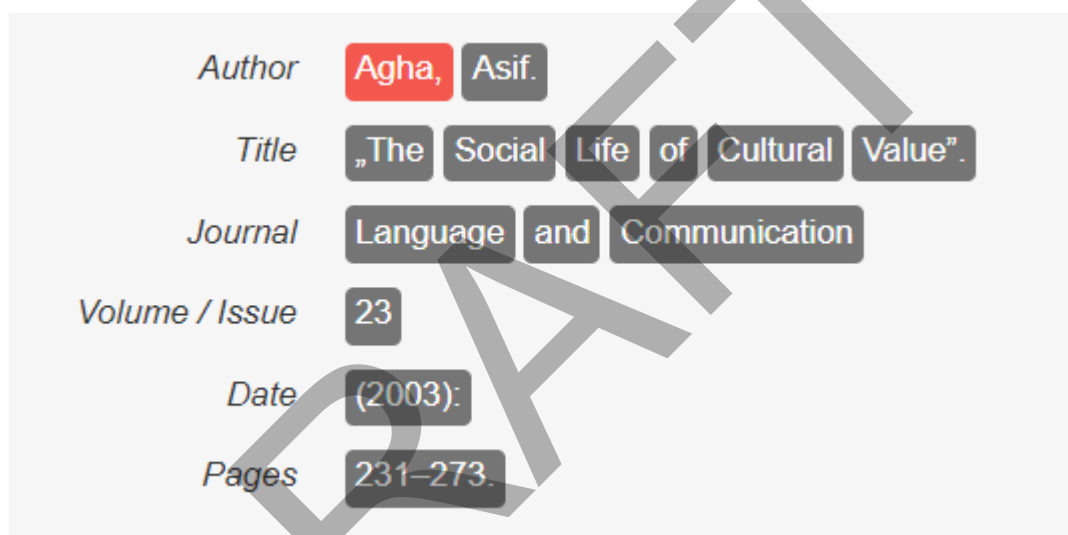


Figure 6.2.3. Example of automatic parsing using Annif software.

As we may observe, the reference enrichment largely depends on the availability and quality of external resources. SSH publications often cite older resources that have not been digitised and in that case a particular datasets would be needed to enrich such citations – printed bibliographies that may not have been digitised. Those resources need to be subjected to the retrospective conversion to obtain properly formatted bibliographic information.

¹⁶⁴ <https://arxiv.org/abs/2205.14677>

6.3. Guidelines for citations as humanities research data in the context of TRIPLE – summary

- 1) Humanities need to **take full advantage of existing infrastructure for open citations** – such as OpenCitations– which includes for example registering abstracts through Crossref services. Through this **humanities data becomes more accessible and easier to link to different data spaces.**
- 2) In the humanities, **reliance on citation recognition and linking based on DOI yields limited results.** What is far more promising are approaches that allow for **citation recognition through more varied reference datasets** (such as national libraries databases or local repositories). Citations are relevant, even without global PIDs present in the bibliographic description.
- 3) In the humanities citations extraction is more challenging as the bibliographic styles and references formatting are more varied. **Dedicated approaches are needed to fully achieve a breakthrough in the humanities citations extraction.** This might be, however, challenging, as many services (such as metadata aggregators or citation indexes) would still rely on DOI-based solutions for citation extraction. A solution tailored for the needs of humanities¹⁶⁵ would call for, for example, reusing the cultural heritage collections – and other local data collections – to ensure citations identification and linking. A novel approach is needed to tackle systemic challenges in this regard. For example, most current solutions for citations extraction and identification rely on using resources like Crossref to recognize a reference and provide their structured expression. For humanities it makes sense to use – besides resources like Crossref – also national libraries catalogues or national databases for research output.

¹⁶⁵G. Colavizza, S. Peroni, M. Romanello, “The case for the Humanities Citation Index (HuCI)”, A citation index by the humanities, for the humanities. International Journal on Digital Libraries, (2022), <https://doi.org/10.1007/s00799-022-00327-0>.

7. Conclusion

GoTriple is an important dataset that is interesting for all those who are interested to understand and analyse European (and wider) societies and cultures and/or would like to enrich their services or datasets through normalised and carefully selected SSH data.

As a research data collection – available via open protocols – it can be reused by a number of stakeholders such as metadata aggregators (knowledge graphs, semantic search engines etc.), indexes and information services (metrics providers, data analysis companies, current research information systems). Its future development will aim to make it more scalable and interoperable to facilitate this kind of interoperability. To leverage this potential it is crucial to understand the key components of the GoTriple data model and the pathways to their improvement. This report offered a discussion on these components and the guidelines for shaping research data collections such as GoTriple.

The main finding and recommendation is that to secure organic development of the service like GoTriple it is **necessary to work on intrinsic modifications** of the current data quality and data model, but it is **equally important to be involved in the larger processes and workflows** that pertain to key elements of the data model and its development (such as PIDs, keywords, persistent identifiers, citations). This report identifies initiatives related to research data quality that are worth engaging into (i.e. Open Abstracts, OpenCitations, Dublin Core-related working groups) or technologies demanding closer and continued inspection (NLP, ML). This also helps to contextualise a heavily discussed issue of specificity of the humanities data – **although we have identified numerous dimensions of this specificity, many crucial developments are discipline agnostic and can help the humanities community face their own challenges if they are properly understood and applied**. One example is implementation of Linked Open Data and Semantic Web technologies in vocabulary creation (which helps in semantic linking of dispersed and heterogeneous ontologies or thesauri). Another is NLP and ML which could drastically impact the ability to produce multilingual content.

In other words, the key to development of the services like GoTriple – services that process humanities research data and aim to provide the output of this processing for further reuse, also scientific and analytical one – **is the creation of environments where humanities (SSH) domain experts, computer scientists and research data experts** can work together in a sustained manner. This deliverable is an example of this kind of effort of bringing together various actors in search for solutions relevant to the challenges which contemporary humanities are facing.

Bibliography

- Aggarwal, C.C. (2018). *Machine Learning for Text*. Springer, Cham.
- Angelaki, G., Badzmierowska, K., Brown, D., Chiquet, V., Colla, J., Finlay-McAlester, J., Grabowska, K. et al. (2019). *How to Facilitate Cooperation between Humanities Researchers and Cultural Heritage Institutions. Guidelines*. Warsaw, Poland: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences. <https://doi.org/10.5281/zenodo.2587481>.
- Balula, A., Caliman, L., Fiorini, S., Jarmelo, S., Leão, D., Mounier, p., Nomine, J.-F. et al. (2021). *Innovative Models of Bibliodiversity in Scholarly Publications: OPERAS Special Interest Group Multilingualism White Paper*. Zenodo. <https://doi.org/10.5281/zenodo.5653084>.
- Bauer, F., Kaltenböck, M. (2012). *Linked Open Data: The Essentials*. Vienna: edition mono/monochrom, 2012. <https://www.reeep.org/linked-open-data-essentials>.
- Boudin, F. (2018). Unsupervised Keyphrase Extraction with Multipartite Graphs. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2*. New Orleans 2018. <https://arxiv.org/abs/1803.08721>.
- Bougouin, A., Boudin, F., Daille, B. (2013). TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. *International Joint Conference on Natural Language Processing (IJCNLP), Nagoya*. <https://hal.archives-ouvertes.fr/hal-00917969/>.
- Britt Holbrook, J., Penders, B., de Rijcke, S. (2019). The humanities do not need a replication drive. *CWTS Blog (archive)*. <https://www.cwts.nl/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive>;
- Car, N., Golodoniuc, P., & Klump, J. (2017). The challenge of ensuring persistency of identifier systems in the world of ever-changing technology. *Data Science Journal*, 16. <http://doi.org/10.5334/dsj-2017-013>.
- Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., Dhillon I. D. (2020). Taming Pretrained Transformers for Extreme Multi-label Text Classification. *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* New York 2020. https://dl.acm.org/doi/abs/10.1145/3394486.3403368?casa_token=-E8vT-FHdnwAAAA:PmQdjkC9gsvvgsZok4T3MptmbPNOzh6DyGC1MkElwOwDnwaV1S0OF5IXKsRVVol7tqQBOjG3Xmc.
- Cioffi, A., Peroni, S. (2022). Structured References from PDF Articles: Assessing the Tools for Bibliographic Reference Extraction and Parsing. <https://doi.org/10.48550/arXiv.2205.14677>.
- Colavizza, G., Peroni, S., & Romanello, M. (2022). The case for the Humanities Citation Index (HuCI): A citation index by the humanities, for the humanities. *International Journal on Digital Libraries*. <https://doi.org/10.1007/s00799-022-00327-0>
- De Santis, Luca. (2022). TRIPLE Deliverable: D6.6 API's Development -RP3 (Draft). Zenodo. <https://doi.org/10.5281/zenodo.7371832>
- Devlin, J., Chang, M-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *NAACL-HLT*, 1.
- Edmond, J., Horsley, N., Kalnins, R., Lehman, J., Priddy, M., Stodulka, T. (2018). *Big Data & Complex Knowledge. Observations and Recommendations for Research from the Knowledge Complexity Project*. University College Dublin. https://kplexproject.files.wordpress.com/2018/04/trinity-big-data-report-jklr_04.pdf.
- Edmond, J., Tóth-Czifra, E. (2018). Open Data for Humanists, A Pragmatic Guide. <https://doi.org/10.5281/zenodo.2657248>.

El-Beltagy, S. R., Rafea, A. (2010). KP-Miner: Participation in SemEval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala. <https://aclanthology.org/S10-1041.pdf>.

European Commission (2020): PID Architecture for the European Open Science Cloud. <https://doi.org/10.2777/525581>.

European Commission. (2021). Towards a reform of the research assessment system: Scoping report. (KI-09-21-484-EN-N). Publications Office. <https://doi.org/10.2777/707440>.

Fellbaum, Ch. (ed.). (1998). *WordNet – An Electronic Lexical Database*, Cambridge, MA: MIT Press.

Ferwerda, E., Pinter, F., Stern, N. (2017). A Landscape Study on Open Access and Monographs: Policies, Funding and Publishing in Eight European Countries. Zenodo. <https://doi.org/10.5281/zenodo.815932>.

Florescu, C., Caragea, C. (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1. <https://aclanthology.org/P17-1102/>.

Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic Recognition of Multi-Word Terms: the Cvalue/NC-value Method. *Int. Journal on Digital Libraries* (3).

Gambhir, M., Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artif Intell Rev* 47.

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., Hiom, D. (2016). A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval'. *Journal of the Association for Information Science and Technology*, 1. <https://doi.org/10.1002/asi.23600>.

Gould, M. (2022). People, places, and things: Persistent identifiers in the scholarly communication landscape. *College & Research Libraries News*, 83(9). <https://crln.acrl.org/index.php/crlnews/article/view/25638>.

Grootendorst, M. (2020). *Keybert: Minimal Keyword Extraction with Bert*. 10.5281/zenodo.4461265.

Gualandi, B., Pareschi, L., Peroni, S. (2022). What Do We Mean by “Data”? A Proposed Classification of Data Types in the Arts and Humanities. arXiv. <https://doi.org/10.48550/arXiv.2205.06764>.

Gupta, S. Gupta, S.K. (2019). Abstractive summarization: An overview of the state of the art, *Expert Systems with Applications*, 121.

Hakala, J. (2010). Persistent identifiers: an overview. *KIM Technology Watch Report*. <http://www.persid.org/downloads/PI-intro-2010-09-22.pdf>.

Haris, G., Błaszczyńska, M., Maryl, M. (2023). ‘TRIPLE Deliverable: D2.4 Report on Identification and Creation of New Vocabularies’, <https://doi.org/10.5281/ZENODO.7539922>.

Harrower, N., Maryl, M., Biro, T., Immenhauser, B., (2020). Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities. Digital Repository of Ireland. <https://repository.dri.ie/catalog/tq582c863>.

Hasan, T., Bhattacharjee, A., Saiful Islam, Md., Mubasshir, K.S., Li, Y.-F., Yong-Bin K., Sohel Rahman, M., Shahriyar, R. (2021). XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics. <https://aclanthology.org/2021.findings-acl.413/>.

Hutchins, B. I., Baker, K. L., Davis, M. T., Diwersy, M. A., Haque, E., Harriman, R. M., Hoppe, T. A., Leicht, S. A., Meyer, P., & Santangelo, G. M. (2019). The NIH Open Citation Collection: A public access, broad coverage resource. *PLOS Biology*, 17(10), e3000385. <https://doi.org/10.1371/journal.pbio.3000385>.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification”. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2. Association for Computational Linguistics., Valencia, Spain 2017.

Khan, A., Salim, N. (2014). A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59.1.

- Klein, M. and Balakireva, L. (2020). On the persistence of persistent identifiers of the Scholarly Web. <https://arxiv.org/abs/2004.03011>
- Kotarski, R., et al. (2021). Developing Identifiers for Heritage Collections. Zenodo. <https://doi.org/10.5281/zenodo.5205757>.
- Krystinski, W., McCann, B., Xiong, C., Socher, R. (2020). Evaluating the Factual Consistency of Abstractive Text Summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Kunze, J., Calvert, S., DeBarry, J.D., Hanlon, M., Jenee, G. and Sweat, S. (2017). Persistence Statements: Describing Digital Stickiness. *Data Science Journal*, 16 1–11, DOI: <https://doi.org/10.5334/dsj-2017-039>.
- Leão, D., Angelaki, M., Bertino, A., Dumouchel, S., Vidal, F.. (2018). OPERAS Multilingualism White Paper. Zenodo. <https://doi.org/10.5281/zenodo.1324026>.
- Lin, H., Ng, V. (2019). Abstractive Summarization: A Survey of the State of the Art. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01).
- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8.
- Marciniak, M., Mykowiecka, A., Rychlik, P. "TermoPL — a flexible tool for terminology extraction". In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*. Portorož, Slovenia., European Language Resources Association (ELRA), European Language Resources Association (ELRA).
- Martín-Martín, A. (2021). Coverage of open citation data approaches parity with Web of Science and Scopus, <https://opencitations.hypotheses.org/1420>.
- Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1). <https://doi.org/10.1007/s11192-020-03690-4>.
- Maryl, M., Błaszczńska, M., Szleszyński, B., Umerle, T. (2021). Dane badawcze w literaturoznawstwie, *Teksty Drugie. Teoria literatury, krytyka, interpretacja*, 2. <https://journals.openedition.org/td/14190>.
- Mihalcea, R., Tarau, P. (2004). TextRank. Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <https://aclanthology.org/W04-3252.pdf>
- Nan, F., Nallapati R., Wang Z., Nogueira dos Santos, C., Zhu, H., Zhang D., McKeown K., Xiang, B. (2021). Entity-level Factual Consistency of Abstractive Text Summarization. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.
- O'Sullivan, J. (2019). The humanities have a 'reproducibility' problem. *Talking Humanities*. <https://talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/>.
- Okulska, I. (2020). Team Up! Cohesive Text Summarization Scoring Sentence Coalitions. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J.M. Zurada, J.M. (eds) *Artificial Intelligence and Soft Computing. ICAISC 2020. Lecture Notes in Computer Science*, vol 12416. Springer, Cham.
- Overton, J.A. et al. (2020). String of PURLs – Frugal Migration and Maintenance of Persistent Identifiers. *Data Science*, 3(1). <http://doi.org/10.3233/DS-190022>.
- Papagiannopoulou E., Tsoumakas G. (2019). A Review of Keyphrase Extraction, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2. <https://arxiv.org/pdf/1905.05044.pdf>.
- Paskin, N. (2010). Digital object identifier (DOI®) system. *Encyclopedia of library and information sciences*, 3, 1586-1592. http://0-www.doi.org.oasis.unisa.ac.za/topics/020210_CSTI.pdf.

- Peels, R. (2019). Replicability and replication in the humanities. *Res Integr Peer Rev* 4. <https://doi.org/10.1186/s41073-018-0060-4>;
- Peels, R., Bouter, L., van Woudenberg, R. (2019). Do the humanities need a replication drive? A debate rages on. *Retraction Watch*. <https://retractionwatch.com/2019/02/13/do-the-humanities-need-a-replication-drive-a-debate-rages-on/>.
- Plomp, E. (2020). Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments. *Data Science Journal*, 19(1). <http://doi.org/10.5334/dsj-2020-046>.
- Pramita Widyassari, A., Rustad, S., Fajar Shidik, G., Noersasongko, E., Syukur, A., Affandy, A., Rosal Ignatius Moses Setiadi, D. (2020). Review of automatic text summarization techniques & methods. *Journal of King Saud University – Computer and Information Sciences*, 4.
- Pęzik P., Mikołajczyk A., Wawrzyński A., Nitoń B., Ogrodniczuk M. (2022). Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer, *ACIIDS2022*. <https://arxiv.org/abs/2209.14008>.
- Sharma, G., Sharma, D. (2022). Automatic Text Summarization Methods: A Comprehensive Review. *SN COMPUT. SCI.* 4.
- Shearer, K., Chan, L., Kuchma, I., Mounier, P. (2020). Fostering Bibliodiversity in Scholarly Communications: A Call for Action. Zenodo. <https://doi.org/10.5281/zenodo.3752923>.
- Shotton, D. (2013). Open citations. *Nature*, 502 (7471). <http://dx.doi.org/10.1038/502295a>.
- Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, 126(6). <https://doi.org/10.1007/s11192-021-03948-5>.
- Spinaci, G., Colavizza, G., & Peroni, S. (2022). A map of Digital Humanities research across bibliographic data sources. *Digital Scholarship in the Humanities*, fqac016. <https://doi.org/10.1093/llc/fqac016>.
- Strader, C. R. (2011). Author-Assigned Keywords versus Library of Congress Subject Headings. *Library Resources & Technical Services*, 4. <https://doi.org/10.5860/lrts.53n4.243>.
- Suominen, O., Inkinen, J., Lehtinen, M. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing'. *JLIS.it*, 1. <https://doi.org/10.4403/jlis.it-12740>.
- Tay, A., Kramer, B., Waltman, L. (2020). Why openly available abstracts are important — overview of the current state of affairs. <https://medium.com/a-academic-librarians-thoughts-on-open-access/why-openly-available-a-bstracts-are-important-overview-of-the-current-state-of-affairs-bb7bde1ed751>.
- The PID Forum. (2019) Why Use Persistent Identifiers? <https://pidforum.org/t/why-use-persistent-identifiers/714>.
- Tóth-Czifra, E., Truan, N. (2021). Creating and Analyzing Multilingual Parliamentary Corpora. <https://halshs.archives-ouvertes.fr/halshs-03366486>.
- Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1). https://doi.org/10.1162/qss_a_00112.
- Vogel, D. (2014). Qualified Dublin Core and the Scholarly Works Application Profile: A Practical Comparison. *Library Philosophy and Practice (e-Journal)*, 1. <https://digitalcommons.unl.edu/libphilprac/1085>.
- Walk, P. (2019). PIDs in Dublin Core. <https://doi.org/10.5281/ZENODO.2551181>.
- Wittenburg, P. (2019). From persistent identifiers to digital objects to make data science more efficient. *Data Intelligence* 1. http://doi.org/10.1162/dint_a_00004.
- Wydmuch, M., Jasinska, K., Kuznetsov, M., Busa-Fekete, R., Dembczyński, K. (2018). A no-regret generalization of hierarchical softmax to extreme multi-label classification. *Advances in Neural Information Processing Systems*, 31.

Appendix: Data Enrichment in TRIPLE: the current state of work

Luca De Santis¹⁶⁶

This section presents the data enrichment strategies devised in the TRIPLE project, focusing on the main topics of the booksprint, that is abstract and keywords management. Also, a small digression about Permanent Identifiers (PIDs) in GoTriple is presented at the end.

All the platform's data are automatically harvested from various sources by using GoTriple's metadata ingestion and curation service, named SCORE.

SCORE processes data by using a pipeline approach: it consists in fact of several specialised components, developed by using the Apache Camel¹⁶⁷ technology, each dedicated to implement a particular feature of a "data flow", starting from the retrieval of the single information (a publication but also a project description), the curation and the enrichment of its metadata and finally the memorisation of the final result in the platform indexes, implemented through the Elasticsearch search engine¹⁶⁸.

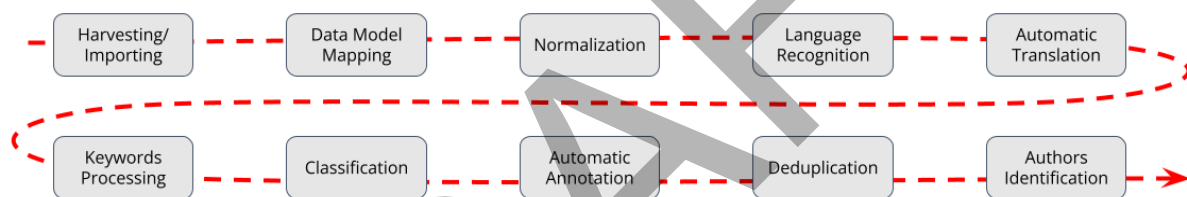


Figure Appendix 1. Publications data flow

At present publications metadata in GoTriple are acquired in three possible ways: by harvesting OAI-PMH Endpoints, with data formatted either in Dublin Core or in the Europeana Data Model, plus by importing data dumps (archives of metadata files in JSON or XML formats) from OpenAIRE and Isidore.

At the time of writing, more than 4.3 million publications and datasets from aggregators and providers, including Isidore, OpenAIRE, EKT, DOAJ, Biblioteka Nauki, ZRC Sazu, Cessda and Coimbra University have been integrated. Soon publications metadata from Open Edition, BASE and Europeana will be also available.

Harvesting data from different sources proved to be a challenge in TRIPLE. Several issues were encountered, including:

- difficulties to map metadata in the TRIPLE data model, especially if they are in a format with a limited expressivity as Dublin Core.
- trying to manage the "personal extensions/interpretations" of the standard made by data providers

¹⁶⁶ Net7, <https://orcid.org/0000-0003-0527-840X>

¹⁶⁷ <https://camel.apache.org/>

¹⁶⁸ <https://www.elastic.co/elasticsearch/>

- dealing with multilingualism for specific attributes (in particular authors when they are mentioned in multiple languages with different alphabets)
- issues in data quality, including plain mistakes (errors in dates or in language attribution)
- using multiple codifications for dates or languages (e.g. “en”, “en-us”, “eng”, ...)
- the frequent use of textual strings instead of standard vocabularies for many attributes.

The general rules that have been followed in the TRIPLE project for data normalisation are:

- removing duplicates when they appear
- cleaning textual strings, by trimming leading and trailing spaces and removing all the HTML codes in them
- defining a controlled vocabulary for some elements
- for normalised attributes, always maintaining the original metadata received, which are therefore stored in separate elements of the final GoTriple publications index on Elasticsearch.

Processing **abstracts** is in general a straightforward task. Normally abstracts are found in a dedicated element with its language attribute. Some typical erroneous situations include:

- sometimes an abstract is originally split in multiple elements, possibly for a wrong interpretation of paragraphs by the data provider,
- language attributes might be missing or wrong,
- text contains HTML elements.

In the first case we take the elements of the same language and merge them in a single abstract.

It has been decided, after doing tests on a sample from the first GoTriple dataset, to always detect the language with the Apache Tika¹⁶⁹ library and store its code in ISO-639-1 format.

Text cleaning includes the removal of HTML and of leading or trailing blank characters.

If there is no English translation the abstract, together with the title of the article, is translated with the eTranslation service¹⁷⁰. This way we can guarantee in GoTriple to practically always have an English version of textual descriptions.

Finally titles and abstracts are used for the automatic Classification (TRIPLE/Moress Categories) and Annotation (TRIPLE Vocabulary) enrichment.

As far **keywords** are concerned, there are two types of them in GoTriple:

- free text keywords from the original sources
- TRIPLE vocabulary entities, detected with the annotation service developed by the French company Foxcub.

We discuss here only the first type, whose normalisation proved quite tricky and gave way to many discussions amongst the TRIPLE consortium.

¹⁶⁹ <https://tika.apache.org/>

¹⁷⁰ <https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>

A basic curation is always applied to keywords, including the removal of duplicates or trimming the blank spaces before and after every string.

It was also decided to normalise the language attribute associated with them if present (the “lang” attribute) by using a controlled vocabulary (see TRIPLE deliverable D2.5¹⁷¹). At the same time, contrary to what has been implemented for titles and abstracts, the language code of keywords is maintained instead of assessing it with Apache Tika.

Another decision that was taken is to remove any possible keyword that refers to codes or labels of taxonomies used by data providers. This was necessary to present data in a cleaner way and to improve search facet filtering in GoTriple. The rule to identify whether a keyword is a taxonomy element or not has been adapted to the various data sources and presented in TRIPLE deliverable D2.5 “Report on data enrichment”.

The “removed” keywords are not lost but they are stored in a special attribute (discarded_keywords) of the Elasticsearch index.

Finally, it has been decided to accept as single keyword strings containing commas or points: no assumption is therefore made whether a case like this refers to a single element or a set of keywords.

Other curation and enrichment done in TRIPLE for publications data include:

- publication date normalisation: we accept only dates after after 1.700 AD; all dates are formatted according to the ISO 8601 format (yyyy, yyyy-mm, yyyy-mm-dd)
- controlled vocabularies have been introduced for:
 - language codes, with 24 ISO-639-1 language codes plus the “other” and “undefined” elements
 - document types, with 18 elements, mapped with the corresponding COAR¹⁷² resource types plus “other” and “undefined”
 - licences, with 11 licences plus “other” and “undefined”
 - access rights, with 7 elements plus “other” and “undefined”
- Publication deduplication
- Authors disambiguation, that is, trying to match as the same person authors names spelled differently (e.g. “Suzanne Dumouchel”, “Dumouchel, Suzanne”, “Dumouchel, S.”).

It is paramount for TRIPLE to guarantee a **wide reuse of its data**. At present this is possible either by using the public search REST APIs or the OAI-PMH endpoint. It is also very valuable to embrace a Linked Open Data approach, in which GoTriple entities are returned in a semantic open data format (e.g. JSON-LD) described by an official TRIPLE ontology that takes into account not only the TRIPLE Data Model but also the controlled vocabularies introduced in the normalization phase cited above. Experiments in this sense are undergoing and their results will be presented at the TRIPLE 2023 conference in Bonn.

¹⁷¹ L. De Santis. TRIPLE Deliverable: D2.5 - Report on Data Enrichment (Draft). Zenodo. 2022. <https://doi.org/10.5281/zenodo.7359654>.

¹⁷² https://vocabularies.coar-repositories.org/resource_types/

As far as **Persistent Identifiers** (PIDs) are concerned, at present this concept isn't managed in GoTriple. There isn't even an internal TRIPLE ID, since documents are identified by the original main identifier of their data sources.

During an informal discussion on this topic two possibilities were considered:

- using GoTriple URL/URI as PID,
- doing an integration with some established service, in particular handle.net.

The first option comes from the Semantic Web experience and it is based on improving the current URL logic in GoTriple by considering three parts:

- the access protocol and the domain: `https://gotriple.eu`,
- the class of the element in singular form (e.g. "document" instead of "documents", as it is now),
- a unique ID generated with a solid random mechanism which guarantees negligible risks to generate duplicates, e.g. Nano ID¹⁷³.

The result is both an ID and an "actionable" URL that can be accessed on the web. Content negotiation mechanism could also distinguish:

- requests of linked data, e.g. done by an application, which return all the document's information in JSON-LD, formatted accordingly to the (yet to be developed) TRIPLE ontology,
- or if the access comes from a browser (or a search engine bot): in this case there should be a 302 HTTP redirection to a SEO-friendly URL, with the added title at the end. This is shown in the image that follows.

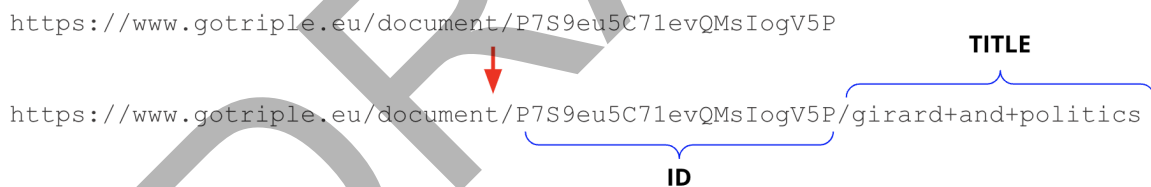


Figure Appendix 2. PID URL redirection to a SEO and user friendly URL

On the other hand, the Handle.net integration might exploit a (yet to be established) OPERAS registry entry. Specific IDs, produced according to the Handle.net formatting rules, would be created for every document and sent to the Handle.net registry by a new SCRE service.

As we are approaching the final part of the project, **introducing changes in metadata management** can become quite complicated, especially regarding PIDs. Every structural change in metadata translates into the reprocessing of all the existing data, which is very time consuming and can take several weeks to finish. The introduction of PIDs in particular might also have an impact on the data collected so far by the Recommender, which are based on the current way used by GoTriple to identify documents.

¹⁷³ <https://zelark.github.io/nano-id-cc/>

At the same time it is important to keep a “long-term” perspective on GoTriple as an OPERAS service, therefore as an infrastructure built to last. As a consequence, all possible improvements must be considered and evaluated, either for being implemented in these few remaining months of the TRIPLE project or in the light of a future evolution of the platform.

DRAFT