# Informetric analyses and support for open science

Due to the exponential growth in the number of scientific works, traditional methods of scientific publications no longer meet the needs in several respects. On the one hand, tools are needed to select those of interest from many publications. On the other hand, in thinking about scientific publications, we rely too much on the frameworks set by the printing. When a publication is published online, it allows direct links to other publications or data. On the other hand, corrections and additions can be dynamic (earlier a special "Errata" contribution was required). Of course, such innovations bring new challenges. The scientific publication system has degenerated anyway. Therefore, it calls for a thorough and well-thought-out overhaul [16]. The answers to the problems are offered by the open science movement.

Open science activities require theoretical and informational support. The latter is provided by informetrics based on bibliographic data. The new open bibliographic database OpenAlex plays a special role in this.

## 23.1. Scientific background, problem identification, and objective of the proposed research

From the beginning, human communities accumulated and transmitted knowledge from generation to generation. The invention of drawing and the writing system based on it (3500-3000 BC) enabled the "storage" of knowledge and its dissemination in space and time. Euclid's Elements represents the first known attempt to summarize and organize the accumulated geometric knowledge based on logic. Most of the knowledge of the Mediterranean and Near Eastern peoples was collected in the Library of Alexandria.

After the collapse of the Roman Empire, Christianity in Europe "censored" the ancient legacy. A good part of it was preserved by the Arabs and enriched with the knowledge of India and Central Asia. Unfortunately, much of what was collected was destroyed by the Mongols during the occupation of the eastern Arab lands (Baghdad Library).

A major problem in the accessibility and durability of accumulated knowledge was the slow and expensive reproduction - copying. Therefore, a very important step in the development of the preservation and dissemination of knowledge was the invention of printing (Gutenberg, 1450). This greatly accelerated and reduced the cost of reproducing works, thereby enabling much wider accessibility and greater permanence of knowledge. Books and other forms of works (leaflets, newspapers, etc.) began to be printed. Scientific societies appeared and began to publish their journals (French, English).

During the Enlightenment, in France, they began publishing an Encyclopedia, which was supposed to summarize all human knowledge. Libraries also expanded and the number of works increased. Over time, the number of works outgrew the memory capabilities of librarians. To review the works collected in the library, a list of works was prepared - often in book form.

The problem with this solution was the introduction of changes. These were written in the appropriate places. As a more appropriate solution, card catalogs (author, subject) - collections of cards began to be used. Each card contained information about an individual work. The cards were stored in drawers in the chosen order (for quick search). The peak of this approach is the Mundaneum catalog (Otlet & La Fontaine, Brussels 1895-1939), which contained 18 million cards in 15 thousand drawers with information about most of the works published up to that time [54].

In his article "As We May Think" (1945), Vannevar Bush suggested a solution (then based on microfilm)

> "Consider a future device ... in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory." [18]

which began to be realized in the 1980s with personal computers.

The development of computers began before World War II (Zuse). They began to enter general use in the 1950s. Initially, large devices were installed in computer centers. Their capacity grew rapidly. In computers, data is stored in digital form (sequences of 0 and 1). Computers have combined media (numbers, text, (moving) images, sound, programs, etc.) in a single device, enabling the storage and processing of large amounts of data. Reproduction (copying) of data in digital form is easy and cheap.

Scientometrics is a subfield of informometrics that studies the quantitative aspects of scientific literature. While the sociology of science focused on the behavior of scientists, scientometrics focused on the analysis of publications. Modern scientometrics is largely based on the work of Derek J. de Soll Price (Science since Babylon, 1961; Little Science, Big Science, 1963) and Eugene Garfield. The latter founded the Institute for Scientific Information (ISI, 1960; now Clarivate Analytics) and created the Science Citation Index, which is often a source of data for scientometric analyses. A dedicated academic journal, Scientometrics, was founded in 1978. The International Society for Scientometrics and Informetrics (ISSI) was founded in 1993 [49, 48].

Informetrics is the study of quantitative aspects of information. It is an extension and refinement of traditional bibliometrics and scientometrics. Informetrics uses bibliometric and scientometric methods to study primarily problems of resource information management and the evaluation of science and technology [49]. Several other related fields also consider other media and forms of scientific materials (notes, reports, plans, patents, specimens, data, programs, recordings, etc.) [35]. The relationships between them are summarized in the figure below by Alexander Doria

Excessive reliance on informetrics in evaluating scientific works is the basis for the "publish or perish" principle, which leads to low-quality research. This principle is one of the manifestations of Goodhart's Law: When a measure becomes a goal, it ceases to be a good measure [36].

Computers can also be connected to each other via data. The development of connectivity began as early as 1970 (ARPANET) and entered wider academic use in the second half of the 1980s (MAIL, FTP, Telnet). General computer connectivity took off in the 1990s in the form of the Internet (World Wide Web (WWW), Tim Berners-Lee, 1990).

Publishing an article in a journal takes time – from a few weeks to a few years. To speed up research in the second half of the last century, authors often printed an article in the form of a report or preprint and sent it to their colleagues. In August 1991, Paul Ginsparg created a preprint repository called **arXiv** on a computer at Los Alamos National Laboratory (LANL), which was accessible from computers connected to the Internet. arXiv is an open-access repository for preprints or post-prints of accepted (non-peer-reviewed) articles in digital

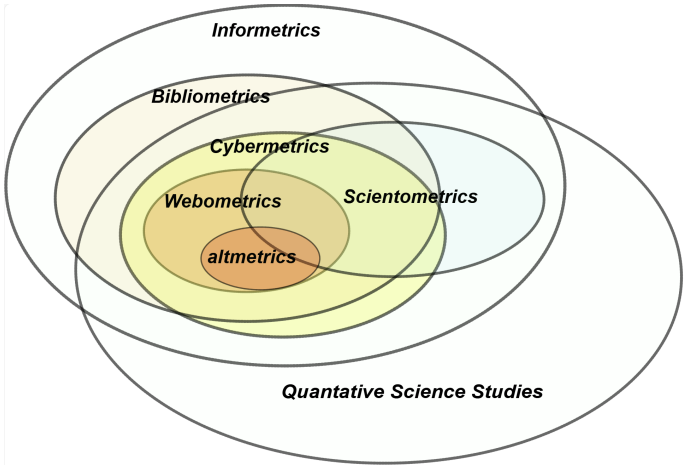


Figure 1: iMetrics

form (e-prints). It accepts articles in mathematics, physics, astronomy, electrical engineering, computer science, quantitative biology, statistics, mathematical finance, and economics. It is imitated by several similar repositories: TechRxiv, ChemRxiv, bioR$\chi$iv, medR$\chi$iv, PsyArXiv, SocArXiv, HAL, RUL, zenodo, etc. Repositories have also emerged for other forms of work: software (c—Net, CPAN, CTAN, CRAN, PyPI (Python), SourceForge), articles (CiteSeerX, ResearchGate, Google Scholar, Academia, DBLP), books (Project Gutenberg, Open Library), data (Kaggle, Microsoft Research Open Data, Google Dataset Search, UCI Machine Learning Repository, CERN Open Data, Dryad, Harvard Dataverse, PEW research, European Social Survey), media (YouTube, Flickr, Google photos, Slideshare, NASA), projects (GitHub, Figshare, Zenodo).

The World Wide Web (WWW) consists of pages with content that can be linked to each other. Each page has its address by which we can reach its content. Initially, the use of the Internet was based on indexes - pages with collections of addresses of interesting pages (usually on a selected topic). Soon the index solution could no longer keep up with the rapid growth of the Internet. It was replaced by collection programs (crawlers) that scan the Internet and add unknown pages to their list. At the user's request, they return the addresses of the pages closest to the request from this list (Lycos, Google). Advertising ("free" service) and the collection of user data (answers and advertisements according to the user's interests) played an important role in improving these search engines. Web 2.0 has enabled (co)creation of online content even by ordinary, non-computer-savvy users. Increasingly, pages are not permanent but are created on the fly at the user's request. [14]

To solve problems on a computer, we need appropriate tools – programs. Most of the time, we have to buy them or write them ourselves. In the late 1960s, Niklaus Wirth developed the Pascal programming language. The Pascal compiler code was available for free. This is one of the earliest examples of free and open software. Openness is very important for the user, because he can (in principle) check what the program does (security) and also find out how (learning). If necessary, he can also adapt such a program to his specific needs (interface language, additional capabilities, etc.). Several such programs appeared, which in the 1980s turned into the GNU free software movement that was formalized by the GNU Manifesto (Richard Stallman, 1985). Free access to collections (libraries, packages) of (sub)programs and also data is very important for the rapid development of science – researchers do not have to start from scratch, but simply build on existing solutions. Towards the end of the 1990s, a related open-source movement emerged,

which also included the possibility of commercial programs. The principle "what was developed with public funds should also be publicly available" has been often applied in American research projects (NSF). This enables the rapid transfer of scientific results into use. A good example of this is the development of the first successful GUI-based web browser, Mosaic, which was the basis for both Netscape and Microsoft's Internet Explorer. Other examples are TEX, R, Python, MySQL, Inkscape, etc.

Google was also created based on a research project. Google showed that with a suitable query, we can quickly get to the relevant content of pages that it has collected on the Internet. We get the answer in the form of a list of hits - web addresses of pages that match the query. Recently, using artificial intelligence, this list has been replaced with a summary that the user can improve in a conversation with the service.

Amazon showed that with descriptions – data about individual things from the real world, these things can be included (made accessible) on the Internet. The online solution can also be used to organize scientific publications. In this case, the online framework is too general. It makes sense to limit ourselves to scientific publications. For an individual work to be accessible, it must be included in the service collection. This is achieved by digitizing works, which has two basic forms:

- full: the work is already created in digital form or has been converted into it, a description of the work is added
- partial: only its description is created for the work, which also contains data on access to the work itself (copyright)

The development of the digitization of historical and cultural resources is the responsibility of the TEI (Text Encoding Initiative). The preparation of online descriptions of works is the responsibility of the Dublin Core Metadata Initiative (DCMI). In the field of librarianship, we have BIBFRAME: Bibliographic Framework Initiative.

Descriptions of works contain various units (author, work, source, institution, country, etc.). It becomes complicated to identify these units when compiling a description. Namely, the data can be ambiguous (different units can have the same name (a string of characters)) or synonyms can occur (the same unit has several names). It is best to solve the problem of identification when entering data into the collection and use unique identifiers for the units (ORCID, DOI, ISSN, ISBN, ISO alpha 2, URL, URI, etc.). This provides high-quality data for informetric analyses. The task could be made much easier by the authors themselves if they used these unique identifiers when preparing their papers – something that journal editors (or paper submission forms) can take care of.

A knowledge graph is a formalism for representing data and knowledge in the form of a graph. Formally, a knowledge graph is defined as a labeled directed graph consisting of a set of triples of the form (Subject, Property, Object). Subject and Object represent the nodes of the graph, and the triples represent the named links of the graph that define the relations between the nodes.

The language for representing knowledge graphs, RDF (Resource Description Framework), was proposed as part of the development of the Semantic Web (Web 3.0). RDF allows for the easy representation of graphs with a text file. Each line of the file contains one triple. For example, the triple (Cankar, Place of Birth, Vrhnika) says that Cankar was born in Vrhnika.

In the last decade, knowledge graphs have been one of the most widely used media for presenting and exchanging data and knowledge in science and industry. In science, RDF is often used to exchange data, such as the results of experiments. In computer science, knowledge graphs are a medium for presenting knowledge from the field of operation of a given information system.

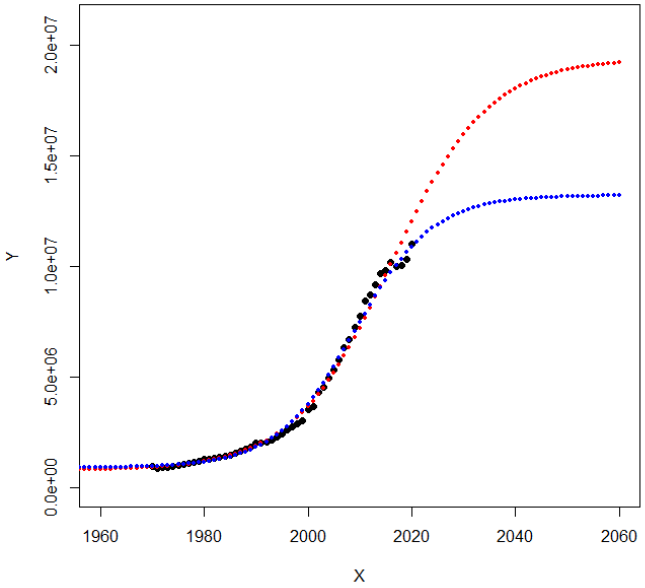


Figure 2: Raised logistic fitting

For example, all major companies, such as Google, Microsoft, Amazon, LinkedIn, Facebook, and others, develop and maintain knowledge graphs for the needs of implementing their information systems.

One of the important characteristics of printed works is their "finality". IT greatly alleviates this. We can correct a digital article, and collect opinions (content comments, ratings) about it, citations to it, the number of hits, etc. For example, we can include 3-dimensional representations [11], very detailed images in vector description [12], or a video clip, which can be viewed with an appropriate viewer. A geography textbook could include tables containing current data about a geographical unit.

From archaeology, we know of records carved in stone, imprinted in clay, carved in wood, written on papyrus, parchment, or paper, etc., which have been preserved in whole or in part for centuries or even millennia and have outlived the civilizations that created them. However, there is a problem with digital records – most recording media (tapes, floppy disks, discs, CDs, DVDs, etc.) have a guaranteed durability of about 5 years [39]. The exception is M-disc discs, which are supposed to last at least 1000 years [51]. An additional problem is access to older reading/writing devices. To ensure the persistence of records, it is recommended to store them in multiple copies on different media and occasionally copy them to newer media. Records are prepared in different formats. Record formats also change. For storage, it makes sense, if possible, to rely on character files and use record formats that preserve the data structure (TeX, SGML, XML, JSON, etc.).

## 23.2. State-of-the-art in the proposed field of research and survey of the relevant literature

Due to the rapid (currently exponential) growth in the number of new works, the current approach to publishing is becoming inadequate (growth in the number of journals, mega-journals, predatory journals, the problem of obtaining reviewers, fraud, etc.).

According to OpenAlex, the number of published scientific papers doubled in the 19 years between 1971 (881,943 papers) and 1989 (1,847,109 papers), then in the next 12 years to 2001 (3,705,036 papers), then in 8 years to 2009 (7,275,504). In 2020, 11,017,156 papers were published. The data for recent years is probably incomplete.

Of course, growth is not unlimited. If we assume a raised logistic model (in the initial part, growth is almost exponential)

$$f(x; L_0, L_1, x_0, k) = L_0 + \frac{L1}{1 + e^{-k\cdot(x-x_0)}}, \qquad L = L_0 + L_1$$

we obtain as local extrema the following fitting curves

blue: $L_0 = 914793.1$, $L_1 = 12310320$, $k = 0.1323116$, $x_0 = 2009$, $L = 13225113$
red: $L_0 = 798003.8$, $L_1 = 18594720$, $k = 0.1064136$, $x_0 = 2016$, $L = 19392724$.

which fit the data quite well, but due to the lack of complete data for recent years, they give very different predictions.

Scopus data from 1996 to 2011 show that 15 million scientists published at least one article during this period, but only 150,608, less than 1%, each year. Their names appear in 41% of all articles and as co-authors of 87% of the most cited articles.

Anomalies and fraud [50]

1. articles with a very large number of co-authors. The article "COVIDSurg Collaborative and GlobalSurg Collaborative: Timing of surgery following SARS-CoV-2 infection: an international prospective cohort study. Anaesthesia 2021, 76, 748–758" has 16162 co-authors.
2. authors who co-authored a very large number of articles in the first ten months of 2024: Wiwanitkit, V. (492), Daungsupawong, H. (346), Bruze, M. (336), etc.
3. degeneration of peer review [4, 34].
4. increasing the importance of institutions by "buying" top scientists [2].
5. predatory journals and conferences [13]: journals with several thousand annual publications (in 2024: Scientific Reports (27528), Heliyon (16271), PLoS ONE (14912), IEEE Access (11726), etc.), thematic issues of journals [52].
6. requests from reviewers or editors to include references to unnecessary sources in the article [37].
7. author fraud (plagiarism, purchased articles, ChatGPT, fabricated/modified data, etc.) [22].

There are indications that there is a change of goals in the informetric monitoring of publications by scientific institutions and funders. Instead of quality, other quantitative parameters are rewarded: e.g. the impact of publications or even the impact of journals in which the articles are published. Users are adapting to this – and the Goodhart law applies. As a result, publishers and journals are multiplying, trying in every way to increase the popularity of publications, regardless of quality. To publish in predatory journals, it is important to have access to public

money. There is a drain of public money into the private pockets of multinational publishing houses.

Ben Torben-Nielsen writes on LinkedIn [45]

> Then there is the money. Authors write papers for free. Reviewers review for free. Editors often work for free. Yet publishers charge readers thousands for access to research. Or they ask authors to pay substantial fees to make their work freely accessible. When selling papers becomes the business model, rigorous peer review becomes an obstacle to profit.

and Daniel Couto comments

> In this day and age Publishers are completely obsolete. No one needs a publisher to make their research known to the public, anything can be uploaded and diffused to the whole world nearly for free. Moreover all the quality that Publishers are meant to ensure is actually being ensured by unpaid reviewers. So why do researchers keep pushing tax payer's money into the pockets of publishing companies? Simply because National funding institutions required publications in such outlets as a metric of success. This metric may have had some merits in the past. Now it brings more harm than good. Sorry, but the system is really broken and makes no sense in todays world.

Open science is a modern approach to scientific research and the dissemination of its results (knowledge) transparently and collaboratively. Opening up science through the fastest possible exchange of information and knowledge between researchers increases the chances for faster progress in science [3]. [43, 24, 1, 28, 32, 17, 21]

In 1942, Robert K. Merton, in his article "A Note on Science and Democracy" (reprinted in the book The Sociology of Science), wrote four "pillars" of science ("communism", universality, work for the common good (selflessness, ethics), skepticism). Later, originality was added [53]. Chubin, D. E. also refers to them in his article "Open Science and Closed Science: Tradeoffs in a Democracy" (1985) [19]. In the age of the Internet, the open science movement has made free access to scientific publications its main pillar, as reflected in the declaration of the Budapest conference in 2001. This is legally regulated by the Creative Commons licenses created in 2002. Subsequently, the requirement for openness has been extended to data and software. Openness can be implemented with online repositories.

For informetric analyses, free access to descriptions (metadata) of published scientific works – including those published in subscription journals – is important. They are striving to make free access also apply to abstracts [46]. An important step in this direction is OpenAlex.

OpenAlex [41] is a completely open catalog of the global research system. It is named after the ancient Library of Alexandria. It was established by the non-profit organization OurResearch. It became accessible in January 2022 via a user interface, a free API, or a snapshot of all the data that can be downloaded to your computer. It is considered a replacement for the Microsoft Academic Graph service, which was discontinued on December 31, 2021. OpenAlex is based on 7 types of units (entities): work, author, resource, institution, concept, publisher or funder. It solves some important issues for the analysis of bibliographic data:

- identification of bibliographic units (ID, resolution)
- free access (sharing of obtained data, transfer to the user's computer)
- content improvement with user participation (the user submits a request for correction)

OpenAlex can be used via a web user interface or programmatically with API calls. For more complex processing, the problem is the limitation to a maximum of one hundred thousand calls per day. This limitation can be avoided by setting up a copy of OpenAlex on a local computer. OpenAlex contains data on most of the works found in the Web of Science and Scopus commercial services. In addition, there is a variety of other works – for example, data on publications in the arXiv repository. This allows us to monitor what are the current "hot" topics in science. Data from commercial services are limited to established journals and contain a time lag caused by the publication process of articles. OpenAlex also contains data on the publication mode (openness) of each work

To use data from OpenAlex, it is necessary to develop appropriate software support that collects and converts the desired data into a format suitable for data (statistical, network, AI) analysis. Support for data cleaning and quality control is particularly important. Sometimes the data also needs to be refined – for example, determining the category of references (agreement, comparison, definition, difference, disagreement, hypothesis, method, position, result, similarity) [15].

The expectations of an author-scientist on scientific publishing are:

1. immediate publication and access,
2. "registration" – timestamp,
3. easy to find,
4. unlimited access,
5. permanent storage (longevity, durability),
6. reader (community) response, acknowledgments.

This can be achieved with repositories of works. The evaluation of the work is not the main concern of the author – it is important for his/her employers, funders, etc.

The evaluation of scientific results places too much emphasis on the number of articles and the number of citations of articles. There are other forms of work: books, patents, plans, films, data, programs, websites, etc. Recently, in a webinar, the famous statistician John Bailer said: what are my few thousand citations compared to the millions using Wickham's Tidyverse. Several movements and recommendations attempt to enable and implement a more comprehensive evaluation of scientific research work, such as ORCID, CRediT (Contributor Roles Taxonomy), Leiden Manifesto for Research Metrics, San Francisco Declaration on Research Assessment (DORA), The Vancouver Recommendations, Coalition for Advancing Research Assessment (CoARA). The possibility of repeating experiments and analyses is becoming important, as advocated by the FAIR (Findability, Accessibility, Interoperability, and Reuse) movement [56, 23].

There are indications that in informetric monitoring of publications, goals are being replaced: instead of quality, other quantitative parameters are being rewarded: e.g. the impact of publications or even the impact of journals in which articles are published. Users are adapting to this. Therefore, publishers and journals are multiplying, trying in every way to increase the popularity of publications, regardless of quality. Access to public money is important for publishing in predatory journals. Public money is flowing into the private pockets of multinational publishing houses. Publishing models are changing radically from subscription to open access. APC (Article Processing Charge) and hybrid publishing models are dominant.

In response to these challenges, new publishing models have emerged, such as "publish, review, curate" (PRC). This model reverses the traditional review-then-publish approach by first publishing the work online and then subjecting it to peer review. The aim of this approach is to increase transparency and accelerate the dissemination of research [20].

Building a knowledge graph for the OpenAlex data environment. The OpenAlex knowledge graph will contain an ontology that includes article types, scientific fields, institutions, and other entities related to articles. The construction of the OpenAlex knowledge graph is possible by connecting to existing knowledge graphs, such as DBpedia, WikiData, Yago, and others. These knowledge graphs mostly represent common-sense knowledge about the world.

Newer big data analytics systems, such as the open-source Apache Spark, allow for the implementation of complex analytics on data that is either in text format (JSON) or in graph format. The analytics can be performed using interactive SQL queries or programs in general-purpose programming languages such as Java, Python, and R.

## 23.3. A detailed description of the work program

### WP1: Project management, coordination and dissemination

Management activities will take place during the entire project and will include administrative work related to work performance, quality assurance, risk assessment and mitigation, compliance with the proposal, and contractual obligations with the Slovenian Research and Innovation Agency. In parallel, coordination activities will include fostering agreement on responsibilities, scheduling and coordinating team meetings, and internal briefing on outcomes of individual activities.

This work package consists of four main tasks. A detailed description of the work in this package and overall project management is available in section 23.5. Project management.

**T1.1 Coordination.** There are 3 partners in the project which have already established long-term cooperation. We will monitor the work on the project in monthly seminars.

**T1.2 Reporting.** Done yearly, as required by the financer (SRA/ARIS). The principal investigator will assign a member of the project to coordinate the collection of achievements in the reporting period and to prepare and submit the annual report. Financial reporting and funds monitoring will be performed by the accounting departments of the partners.

**T1.3 Dissemination.** The results will be reported at international conferences and published in scientific journals. The developed software, its documentation, and example data sets will be made available on GitHub or other repositories as open-source. At the beginning of the project, a special project web page will be established.

We will prepare some popularization activities about open publishing (seminars, and articles in some professional journals).

Some project results will be included in our teaching (network analysis, databases). The local version of OpenAlex will also provide a "playground" for experimenting with database systems and a rich and complex source of data for education (seminars, diplomas, master's and PhD works).

**T1.4 Data management.** A data management plan (DMP) will be prepared in the first six months of the project (D1.1.) following the FAIR principles of making data findable, accessible, interoperable, and reusable. The DMP will detail the purpose of data collection, relevance to the objectives of the project, specify the types and formats of data and metadata to be collected, the expected size, and how data will be curated and preserved.

### WP2: Bibliographic data and OpenAlex

**T2.1 OpenAlex** We already developed an R package OpenAlex2Pajek that supports the construction of bibliographic networks using the OpenAlex API [6]. It is included in the OpenAlex support [42]. To enable complex analyses of bibliographic data we will establish a local copy of OpenAlex's bibliographic base and develop software support for its exploitation. We will also consider options for combining OpenAlex data with data from other (mostly Slovenian) bibliographic databases.

**T2.2 Development of OpenAlex usage support** We will study approaches to using big data systems for data analytics and graph database systems to analyze informetric data and try to develop the corresponding methods. To be able to analyze the OpenAlex data in its native exchange format JSON, the data processing environment will be created by using Apache Spark, a data processing engine for large-scale data analytics.

To process the OpenAlex data in a graph form we will convert JSON representation of OpenAlex into RDF format used as part of the Semantic Web (Web 3.0). RDF representation of OpenAlex can be published on Web 3.0 as a knowledge graph, or stored in a graph database system where it can be analyzed.

**T2.3 Compilation of bibliographic data.** Acquisition, cleaning, and preparation of bibliographic data (from OpenAlex and COBISS and administrative data from the University of Primorska and/or other institutions) for analyses in WP3, WP4, and WP5.

### WP3: Informetrics

An important tool in the analysis of collections of linked networks (bibliographic networks are a special case) is network multiplication [9] which enables us to compute derived networks. To consider each unit equally in the analysis of bibliographic networks, the fractional approach is used. Its theoretical background was proposed in the paper [5]. To get efficient algorithms for some problems we consider in "expensive" parts of processing only the relevant results – the truncation approach [10]. We will continue to explore the possibilities provided by these three approaches in the bibliographic network analysis. The main tasks in this WP include:

**T3.1 Theoretical development.** We will continue our theoretical work on the analysis of derived networks and the development of efficient algorithms for it. We will try to provide theoretical and algorithmic support for problems encountered by other WPs. For example, the construction of measures of publication quality or procedures for the detection of non-ethical publication practices.

**T3.2 Open publishing and PRC support.** We will collect and if needed develop tolls for the support of open publishing and the PRC approach such as automatic review procedures: author recognition and connection to ORCID, adding DOI to individual sources, checking language or plagiarism, assessment of the connection of sources with the topic of the article, assessment of the quality of an individual source, selection of reviewers, the relevance of keywords, analysis of the connection between authors and reviewers, identification of unethical activities, related articles, etc.

**T3.3 Application of developed methods for other WPs.** An example, is the problem of

misclassification of journals and researchers. In many places, more than one classification of a research field appears for a unit. (e.g. SICRIS, SCIMAGO). Due to different citation cultures, serious errors can occur in ranking within a single field. We will develop tools to find such anomalies.

Another example is the problem of distinguishing potentially predatory methods in journals with a high impact factor. It is not possible to determine from the impact factor alone whether a journal is suitable for publication. We will define methods that will identify potentially controversial journals with the help of other data.

WP4: Support and monitoring of Open Science

**T4.1 Overview of the state of open science in Slovenia.** This task aims to provide a comprehensive overview of the current landscape of Open Science (OS) practices, policies, and initiatives in Slovenia. It involves conducting a thorough analysis of national policies and frameworks, research institutions and stakeholders, open access and data repositories, technological infrastructure, barriers and challenges, as well as growth opportunities [30, 31, 32, 33].

**T4.2 Support and monitoring of Open Science in Slovenia.** Based on the data gathered and software developed in WP2 and WP3, we will evaluate the completeness of information from different sources and analyze the share of open access (OA) publications throughout the years. Moreover, we will compare different types of OA (Diamond, Gold, and Green) and examine the uptake of OA publishing across different research fields. The analysis will also extend beyond journal articles and include diverse publication types, from monographs to data publications.

**T4.3 Survey on the uptake of Open Science practices and attitudes among researchers.** Based on a review of published results of surveys focusing on the publication behavior of researchers, we will prepare a survey questionnaire to be disseminated among researchers in Slovenia. This task aims to survey researchers to assess the adoption of OS practices and explore their attitudes towards OS, including its perceived benefits, challenges, and impact on researcher evaluation. Based on the survey responses, we will analyze the extent of OS adoption, examine attitudes toward OS, and explore the role of researcher evaluation in influencing OS uptake. The analysis will identify challenges and opportunities for improving OS practices, to provide actionable recommendations for researchers, institutions, and policy-makers to foster a more open and collaborative research environment.

**T4.4 Survey of editors of scientific journals.** This task will involve surveying editors from scientific journals to assess their engagement with OS practices, including OA publishing, data sharing policies, and transparency in the editorial and peer review processes. The survey will explore editors' attitudes toward OS, the challenges they face in implementing open practices, and the support they need to enhance OA publishing models. The results will provide insights into the barriers and opportunities for advancing OS in scholarly publishing and inform recommendations for improving collaboration between editors, researchers, and institutions.

WP5: Scientific publishing.

**T5.1 State of scientific publishing in Slovenia.** We will try to assess the speed of change in the field of scientific publishing, and the increase in numbers in recent years (no. of publications,

no. of journals, APC + subscriptions). Acquisition from WP2, cleaning, and analysis of data on publishing (i.e. data on APCs, editorial process, open peer reviews, PRC, etc.). Analysis (using WP3). Identification of problems, and recommendations.

**T5.2 Problems in publishing.** We will make an overview of the problems in scientific publishing and the proposed approaches how to deal with them. A special focus will be on Goodhart's law (formal description of Goodhart's law; application of Goodhart's law in informetrics; methods for detecting potentially unethical behavior in scientific publications (authors, editors, reviewers, publishers); preparation of recommendations for correction of the situation.)

**T5.3 Open publishing**. We will make an overview of the PRC approach, evaluate its applicability in Slovenian scientific publishing, and prepare recommendations. Some proposed solutions will be tested in editing the open journal Ars Mathematica Contemporanea (AMC) and ADAM (The Art of Discrete and Applied Mathematics). We will try to connect with international open publishing initiatives such as Free Journal Network (FJN) which unites some diamond open-access journals.

## 23.4. Available research equipment over 5.000 €

For most data sets a better laptop with at least 32 GB of memory is sufficient. For very large data sets we will use the computing facilities available at the UP.

## 23.5. Project management: Detailed implementation plan and timetable

The project is spanned over three organizations: University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technology (UP FAMNIT); Institute of Mathematics and Physics, Ljubljana (IMFM); and University of Ljubljana, Faculty of Information Studies in Novo mesto (FIŠ). Roughly, Faculty of Information Studies in Novo mesto and partially UP FAMNIT will be focused mainly on social and information science aspects, while UP FAMNIT and IMFM will be responsible primarily for theory, algorithm design, programming, data processing, and prototyping.

Project management. The principal investigator (PI) will be responsible for the achievement of milestones (reports) and deliverables. He will also ensure that the project is up and running within the set deadlines. Members of work package groups will have weekly meetings to assess progress on the project and identify possible risks. Results and work in progress will be presented in seminars. The communication on the project will be carried out via electronic media and meetings. Documents, data, and computer code will be stored in GIT and other open repositories.

General methodology. The development of methods in data analysis consists of analyzing data, setting approximate conjectures, developing and implementing analytic algorithms supporting the method, testing, and at the end evaluating of results and the method itself. Software development of a library consists of cycles that include planning, implementation, testing, optimization, and documentation. Initial cycles can be simplified and consist only of planning, prototyping, and evaluation for the development of methods that lead consolidation cycles (planned code rewrites). All the above-stated methods are iterative processes.

Project timeline. The expected project duration is 3 years. All work packages will be active throughout the project. Here is the list of planned activities – (from-to) in months

**M2.1.1.** Local copy of the OpenAlex data providing basic access (1-6).
**M2.2.1.** Development of data processing environment for analyzing OpenAlex data (7-36).
**M2.3.1.** Building KG implemented in RDF with the translation from the JSON representation (12-30).
**M2.3.2.** Final KG including higher levels of ontology and merged with DBpedia (31-36).
**M3.1.1.** Work on detection of non-ethical practices (1-36).
**M3.1.2.** Providing theoretical support for other WPs (7-36).
**M3.1.3.** Work on informetric network analysis algorithms (1-36).
**M3.2.1.** Review of open publishing practices, standards and measuring (1-12).
**M3.2.2.** Review of non-ethical practices (1-12).
**M3.2.3.** Development and implementation of algorithms for M3.2.1. and M3.2.2. (13-36).
**M3.3.1.** Application of developed methods for other WPs (7-36).
**M4.1.1.** Review of policies (1-6)
**M4.2.1.** Analysis report and identification of case studies for T4.3 and T4.4 (6-12)
**M4.3.1.** Review of previous research on publishers and questionnaire developoment (6-12)
**M4.3.2.** Data collection and preparation (12-18)
**M4 3.3.** Data analysis and repoting (18-24)
**M4.3.1.** Review of previous research on publishers and questionnaire developoment (12-18)
**M4.3.2.** Data collection and preparation (18-24)
**M4 3.3.** Data analysis and repoting (24-30)
**M5.1.1.** Completing the data collection of stratified APC (1-12).
**M5.1.2.** Completing the analysis (13-18).
**M5.1.3.** Compiling and publishing Recommendations (19-30).
**M5.2.1.** Delivering a presentation on the Goodhart Law on a conference.(1-12).
**M5.2.2.** Submitting a scientific paper on the Goodhart Law (13-24).
**M5.2.3.** Completing a collection of algorithms for detecting anomalies in scientific publications (13-36).
**M5.3.1.** Delivering a presentation on PRC models. (1-12).
**M5.3.2.** Submitting a scientific paper on PRC models (13-24).
**M5.3.3.** Delivering a report on tests performed on AMC and ADAM (13-36).

| | YEAR 1 | | | | | | | | | | | | YEAR 2 | | | | | | | | | | | | YEAR 3 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **WP1: Project management, coordination and dissemination** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T1.1 Coordination | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T1.2 Reporting | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T1.3 Dissemination | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T1.4 Data management | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **WP2: Bibliographic data and OpenAlex** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T2.1 OpenAlex | | | | | | M | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T2.2 Development of OpenAlex usage support | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | M |
| T2.3 Compilation of bibliographic data | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | M | | | | | | M |
| **WP3: Infometrics** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T3.1 Theoretical development | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | M |
| T3.2 Open publishing and PRC support | | | | | | | | | | | | M | | | | | | | | | | | | | | | | | | | | | | | | M |
| T3.3 Application of developed methods for other WPs | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | M |
| **WP4: Support and monitoring of OpenScience** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T4.1 Overview of the State of OS in Slovenia | | | | | | M | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T4.2 Support and monitoring of OS in Slovenia | | | | | | | | | | | | M | | | | | | | | | | | | | | | | | | | | | | | | |
| T4.3 Survey on scientific publishers | | | | | | | | | | | | M | | | | | | M | | | | | | M | | | | | | | | | | | | |
| T4.4 Survey on OS among researchers | | | | | | | | | | | | | | | | | | M | | | | | | M | | | | | | M | | | | | | |
| **WP5: Scientific publishing** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T5.1 State of scientific publishing in Slovenia | | | | | | | | | | | | M | | | | | | | | | | | | M | | | | | | M | | | | | | |
| T5.2 Problems in publishing | | | | | | | | | | | | M | | | | | | | | | | | | M | | | | | | | | | | | | M |
| T5.3 Open publishing | | | | | | | | | | | | M | | | | | | | | | | | | M | | | | | | | | | | | | M |

WP duration plan  Objective duration plan  M Milestone planned

## References


[1] Ahmed, Abubakari; Al-Khatib, Aceil; Boum II, Yap; Debat, Humberto; Dunkelberg, Alonso Gurmendi; Hinchliffe, Lisa Janicke; Jarrad, Frith; Mastroianni, Adam; Mineault, Patrick; Pennington, Charlotte R.; Pruszynski, J. Andrew: The future of academic publishing. Nature Human Behaviour volume 7, pages 1021–1026 13 July 2023. PDF

[2] Ansede, Manuel: Dozens of the world's most cited scientists stop falsely claiming to work in Saudi Arabia. El Pais, Dec 05, 2024. WWW

[3] ARIS: Odprta znanost. WWW

[4] Bal, Mieke: Let's Abolish the Peer-Review System. 2018. WWW

[5] Batagelj, Vladimir. On fractional approach to analysis of linked networks. Scientometrics. May 2020, vol. 123, iss. 2, p. 621-633. DOI:[10.1007/s11192-020-03383-y, WWW

[6] Batagelj, Vladimir : OpenAlex2Pajek – an R Package for converting OpenAlex bibliographic data into Pajek networks. COLLNET 2024, Strasbourg, December 12-14. V: Jain, Praveen Kumar (ur.), et al. Innovations in webometrics, informetrics, and scientometrics: AI-driven approaches and insights. Delhi: Bookwell, cop. 2024. p. 66-77. ISBN 978-93-86578-65-5. PDF

[7] Batagelj, Vladimir, Maltseva, Daria. Temporal bibliographic networks. Journal of informetrics: an international journal. Feb. 2020, vol. 14, iss. 1, art. 101006 (14 p). DOI

[8] Batagelj, Vladimir, Ferligoj, Anuška, Squazzoni, Flaminio. The emergence of a field: a network analysis of research on peer review. Scientometrics. 2017, vol. 113, iss. 1, str. 503-532, DOI

[9] Batagelj, Vladimir, Cerinšek, Monika. On bibliographic networks. Scientometrics. 2013, vol. 96, iss. 3, p. 845-864. DOI

[10] Batagelj, V. Weighted degrees and truncated derived bibliographic networks. Scientometrics (2024). DOI; Online

[11] Batagelj, Vladimir: European Airports core of order 13. X3D

[12] Batagelj, Vladimir: Slovenski politiki - matrični prikaz / vse relacije; obdobje do april 2022. WWW/SVG

[13] Beall's List of Potential Predatory Journals and Publishers. WWW

[14] Berners-Lee, Tim, Hendler, James: Publishing on the semantic web. Nature volume 410, pages1023–1024 (2001). WWW

[15] Bertin, M., Atanassova, I. Linguistic perspectives in deciphering citation function classification. Scientometrics 129, 6301–6313 (2024). DOI

[16] Boulton, Geoffrey; Koley, Moumita: More is not better: the developing crisis of scientific publishing. July 2, 2024. WWW

[17] Brembs, Björn; Huneman, Philippe; Schönbrodt, Felix; Nilsonne, Gustav; Susi, Toma; Siems, Renke; Perakakis, Pandelis; Trachana, Varvara; Ma, Lai; Rodriguez-Cuadrado, Sara. (2021). Replacing academic journals, version 5. DOI; Zenodo

[18] Bush, Vannevar: As We May Think. The Atlantic, July 1945 PDF

[19] Chubin, D. E. (1985). Open science and closed science: Tradeoffs in a democracy. Science, Technology, and Human Values, 10(2), 73-80. PDF

[20] Corker, K. S., Waltman, L., & Coates, J. A. (2024). Understanding the Publish-Review-Curate (PRC) model of scholarly communication. MetaArXiv

[21] Council of the European Union: High-quality, transparent, open, trustworthy and equitable scholarly publishing (approved on 23 May 2023). PDF

[22] Day, Adam; Cucen, Ebru: Searching for Research Fraud in OpenAlex with Graph Data Science. 2022. WWW

[23] FAIR Principles. WWW

[24] Guédon, Jean-Claude: In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing. ARL 2001. PDF

[25] Hardt, Moritz; Megiddo, Nimrod; Papadimitriou, Christos; Wootters, Mary: Strategic classification. Association for Computing Machinery (ACM), New York, 2016, 111–122. ISBN: 978-1-4503-4057-1. MR3629816.

[26] Hennessy, Christopher A.; Goodhart, Charles A. E.: Goodhart's law and machine learning: a structural perspective. Internat. Econom. Rev. 64 (2023), no. 3, 1075–1086. MR4624804.

[27] Horbach, S.P.J.M.(.S., Halffman, W.(.W. The changing forms and expectations of peer review. Res Integr Peer Rev 3, 8 (2018). DOI. PDF

[28] International Science Council: The future of scientific publishing. WWW

[29] Kochetkov, D. (2024, March 21). Evolution of Peer Review in Scientific Communication. DOI; SocArXiv

[30] Krapež, Katarina. Impact of publisher's commercial or non-profit orientation on editorial practices: moving towards a more strategic approach to supporting editorial staff. Learned publishing. [Spletna izd.]. 2023, vol. 36, iss. 4, str. 543–553. ISSN 1741-4857. DOI: 10.1002/leap.1575.

[31] Krapež, Katarina. Editors' responsibility for publishing high-quality research results: a worldwide study into current challenges in quality assessment processes. Lexonomica : revija za pravo in ekonomijo. [Tiskana izd.]. 2022, vol. 14, no. 1, str. 127-152. ISSN 1855-7147. dLib.si, DOI: 10.18690/lexonomica.14.1.127-152.2022.

[32] Krapež, Katarina. Advancing self-evaluative and self-regulatory mechanisms of scholarly journals: editors' perspectives on what needs to be improved in the editorial process. Publications / MDPI. [Online ed.]. 2022, vol. 10, iss. 1, str. 1-18. ISSN 2304-6775. PDF, DOI: 10.3390/publications10010012.

[33] Krapež, Katarina. Pravica do sekundarnega publiciranja znanstvenih del: primerjalnopravna analiza ureditev v Sloveniji in državah članicah EU. Pravnik: revija za pravno teorijo in prakso. 2021, letn. 78, št. 11/12, str. 585-615, 655-656. ISSN 0032-6976.

[34] Levy, Neil: Philosophy, Bullshit, and Peer Review. Cambridge (UK): Cambridge University Press; 2023 Dec. ISSN: 2398-0567; 2514-3832. WWW, PDF

[35] Maltseva, Daria, Batagelj, Vladimir. iMetrics: the development of the discipline with many names. Scientometrics. Oct. 2020, iss. 1, vol. 125, p. 313-359, DOI: 10.1007/s11192-020-03604-4,

[36] Pisanski, T., Batagelj, V., Pisanski, J.: Open Science and Goodhart's Law. IS2024-Cognitive Science, Ljubljana, October 11 2024. p. 20-23. DOI

[37] Marquez, Ronald: LinkedIn: What should #Editors do? WWW

[38] Matt, Christian; Hoerndlein, Christian; Hess, Thomas: Let the crowd be my peers? How researchers assess the prospects of social peer review. 2017. PDF

[39] Mediafix: Durability of storage media. WWW

[40] Nooy, Wouter De, Mrvar, Andrej, Batagelj, Vladimir. Exploratory social network analysis with Pajek. 3rd ed., revised and expanded ed. for updated software. Cambridge [etc.]: Cambridge University Press, 2018. XXX, 455 p, ilustr. Structural analysis in the social sciences, 46. ISBN 978-1-108-47414-6, ISBN 978-1-108-46227-3. DOI

[41] OpenAlex. OurResearch.WWW, January 2025

[42] OpenAlex: Client Libraries. OpenAlex2Pajek. WWW, January 2025

[43] Ouvrir la science! WWW

[44] Stokstad E. The 1% of scientific publishing: Only a handful of researchers manage to publish one or more papers per year. Science. 2014 Jul;11. WWW

[45] Torben-Nielsen, Ben: A confession: as a ex-researcher, I have seen the dark side of peer review. LinkedIn, January 23, 2025.

[46] The Initiative for Open Abstracts. WWW

[47] Wikipedia: arXiv. WWW

[48] Wikipedia: Derek J. de Solla Price. WWW

[49] Wikipedia: Informetrics. WWW

[50] Wikipedia: List of scientific misconduct incidents. WWW

[51] Wikipedia: M-disc. WWW; Verbatim

[52] Wikipedia: Mega journal. WWW

[53] Wikipedia: Mertonian norms. WWW

[54] Wikipedia: Mundaneum. WWW

[55] Wikipedia: Open science. WWW WWW

[56] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI, [WWWWWW