A Scientific Knowledge Graph with Community Detection and Routes of Search. Testing "GRAPHYP" as a Toolkit for Resilient Upgrade of Scholarly Content

Renaud Fabre, Otmane Azeroual, Patrice Bellot, Joachim Schöpfel, Daniel Egret

A SCIENTIFIC KNOWLEDGE GRAPH WITH COMMUNITY DETECTION AND ROUTES OF SEARCH

Testing "GRAPHYP" as a toolkit for resilient upgrade of scholarly content

Renaud Fabre

ORCID 0000-0003-4170-324X

Dionysian Economics Lab (LED), Université Paris 8, Saint-Denis, France

Otmane Azeroual

ORCID 0000-0002-5225-389X

German Center for Higher Education Research and Science Studies (DZHW), Berlin, Germany, (azeroual@dzhw.eu)

Patrice Bellot

ORCID 0000-0001-8698-5055

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Joachim Schöpfel

ORCID 0000-0002-4000-807X

Univ. Lille, ULR 4073 - GERiiCO - Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, F-59000 Lille, France

Daniel Egret

ORCID 0000-0003-1605-7047

Observatoire de Paris, Université Paris Sciences & Lettres, 75006 Paris, France

ABSTRACT

Unlimited change in scientific terminology challenges integrity in scientific knowledge graph (SKG) representation, while current data and modeling standards, mostly document oriented, hardly allow a resilient semantic upgrade of scholarly content. Moreover, results of a "multimodal knowledge acquisition" are required for an efficient upgrade of search methods: « vital nodes » differ among users of the same keyword, due to distinct needs of scientific communities, rooted in their own interpretations and controversies.

Modeling and data are challenged to propose new outcomes, mixing automated information and human choices allowing dynamic community detection: to fulfill this program with GRAPHYP **toolkit**, we identify a workflow ensuring the objectives of integrity and completeness of search management activities. It encompasses **data standards** for « routes » of search, **modeling** of **community detection** and navigation inside SK bipartite hypergraph, and a first **test** with extraction of characteristics of communities' preferences from readings of scholarly content. "Search is not Research" and therefore further work should explore the links between modeling and data recording research contents and "search and select" results in SKG data structure.

Keywords: scientific knowledge graph, routes of knowledge, bi-partite crown hypergraph, parsing of scientific publication, impact assessment and modeling of search results, graph convolutional networks.

INTRODUCTION

Towards "multi-modal scholarly knowledge acquisition"

Even without speaking of "academic anarchy" (Feyerabend, 1975), scientific knowledge is not constructed in a linear and unique way but by competing interests, theories, models and communities (Kuhn, 1962). So, how can scientific knowledge graphs (SKG) represent this reality – the existence of different and competing answers to similar queries? How are SKGs equipped to secure both the integrity and completeness of answers in semantic upgrading of scholarly content?

From those perspectives, "vital nodes" in SKG are still part of a hidden treasure: as the conclusions of Lü et al. (2016) point out, search activities have still to tackle with "how to identify the most influential node or the most influential set of nodes at a finite time t instead of the steady state at $t \rightarrow \infty$, in particular for the continuous dynamical processes ». Indeed, a challenge to SKGs accurate representation, is to adjust their data structure to the continuous evolution of scientific terminology. Current data standards and modeling hardly allow a resilient semantic upgrade of scholarly content: the search for the « vital node » differs with current research trends in user communities, among users of the same keyword, with the course of their controversies and the changing paths of their discoveries.

New modeling and clarified data standards would be appropriate to propose an alternative data structure combining the automated information of the users[1] with functions of assistance to the selections of various communities among the items proposed to their choices, and providing differentiated answers to a same keyword.

Such a functional shift in SKG design implies a systemic change, applied altogether to data standards, to modeling and to testing. In an attempt to give answers to these three steps, we describe below the GRAPHYP toolkit, which aims to represent an integrated workflow for resilient SKG upgrade and use of scholarly content, optimizing the achievement of scientific impact assessment: it proposes a way to model the uses of scientific texts as recorded from search sessions. In the same time we notice that knowledge distillation techniques that aim to "model joint nuances and facets of texts" from pairwise inputs (Chen et al., 2020) are open "to modify proposed architecture to better fit non textual features" and are still confronted to high computational costs.
Current evolutions in data standards, modeling and testing give its context to our approach to "multimodal scholarly knowledge acquisition".

- Data standards: wider description of scholarly contents

Data included in SKG contents, built from publications (research data, articles), currently diversify their sources and types. In addition to uses of bibliometric references and citations, new types of data appear in discussions about the accuracy of standards and measures of « vital nodes » in content networks (Lü et al., 2016). For instance, the tables published in annual reports[2] of global STM Association of Publishers reflect a growing attention to emerging data structures that might be called "routes" of community choices of search practiced by readers: new services on consultation platforms allow the user to be informed about any item, both on its content (article, data…) but also on its user-generated content (citation, download, comment). Meanwhile, SKGs begin to provide community detection: linking users and items provides data on preferences and choices, not just the documents themselves. Data on preferences of SKG users produce annotations that can thus interfere in the upgrading of scientific knowledge. Properly identified, recorded and linked, these data outline the imprint of profiled

[1] https://core.ac.uk/display/225542823?source=2
[2] https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf (pp. 57-8, 72)

meaningful user practices: it supplies a new "added value" in knowledge building. We refer to these evolutions of data standards in Section 1.

- Modeling: an approach to community structure of users and to a "semantic representation of communicated scholarly knowledge"

It is currently reported that pandemic events highlight the role of SKGs in their function of linking multiple sources of information issued from heterogeneous categories of documents (Kejriwal, 2020). Recently too, similar ideas led to a kind of a global systemic approach to data and impact of scientific publication (Aryani et al., 2020) that is said to be developed "beyond journal metrics".

However, a recent article (Jaradeh et al., 2019) on research impact assessment issues still considers digital scholarly contents as "mere analogues of their print relatives", while the same article claims for "first steps of a larger research and development agenda that aims to enhance document-based scholarly communication with semantic representations of communicated scholarly knowledge » towards a « multimodal scholarly knowledge acquisition », leading to an « Open Research Knowledge Graph ». Nevertheless, until recently, sharing information on annotations and other preferences remained difficult because even though graph convolutional networks (GCN) « acquire node representations mainly through aggregating their neighbor information », GCNs were « largely ignoring the community structure" (see Section 2). Changes seem to appear as "the neighborhood information of nodes is aggregated by a graph convolutional network » (Liu et al., 2020). Our modeling is inspired by that framework which could open up to new generations of "interpretable recommendation systems", exploiting the technology of feature extraction with heterogeneous data sets. In such configurations the recommender system provides a "knowledge mapping which enables users to understand the value of recommendations" so that the mechanism and the results of the recommendation both become understandable to the user (Shao et al., 2021).

- Testing: accurate community detection with large data corpuses on readings of scholarly content

The evolutions of data standards and the modeling of SKG towards community detection, in turn require, for an efficient and sharp test, to accurately assess the clustering of user preferences and therefore identify communities of users: data on the reading of scholarly content can be processed on a large scale as a direct indicator of the preferences of identified circles inside the scientific community.

Data at large scale exists to identify communities. The total worldwide download of scientific material (estimated between 2.5 and 3 billion/year), and its recording by parsing user logs, provide access to large corpuses of data on reading practices, which are currently covered by their own metrics, with the "Usage Factor for Journals (UFJ1)"[3]. As observed in the STM Publishers annual report 2018 quoted above: « The consensus view seems to be that downloads (as a proxy for readings) *is a potentially useful complement to citation data but that it should not be seen to replace it, because they reflect different aspects of "using" a research paper* ». The analysis of reading behavior is currently studied (Thelwall, 2020). Testing relevant identification of communities and suitability of GRAPHYP workflow to record their preferences, is the last part of our toolkit presentation (Section 3).

In Section 1, we describe the data standards and show how our approach splits data between searchable "Maps" and multimodal search "Routes". Section 2 presents modeling of communities of search with

[3] This standard is « *the Median Value in a set of ordered full-text article usage data (i.e. the number of successful full text article requests) for a specified Usage Period of articles published in a journal during a specified Publication Period* ».

three main functions of GRAPHYP in assessing the impact of research (clustering, neighboring, scaling) and expressing the preferences of communities of users of scholarly content. In Section 3, we tackle with testing the workflow of GRAPHYP: it analyses large data sets on user practices in reading scholarly content, which offers strong guarantees of integrity in the clustering of community detection. Section 4 presents the directions of further works, which develop our current tracks on "Search and select data processing" within the framework of a global modeling of data on annotations interfacing search practices and updating of content.

Section 1 DATA STANDARDS FOR COMMUNITY DETECTION: SEARCHABLE MAPS AND MULTIMODAL ROUTES OF SEARCH

Knowledge "Maps" can be described as representing the knowledge landscape with its scholarly content, while "Routes" record the variety of travels of Map users. The preferences of users of scientific content (articles, databases, etc.) are recorded using specific data structures, hereafter defined as « Routes » that the communities of researchers use to build, by distinction from « Maps » on which the routes are selected. This distinction is already at work because the current SKG structures are open to several representations of user preferences, which are recorded from their uses of scientific corpuses.

1. "Searchable" Maps: a thesaurus of all communities

In the modeling of Scientific Knowledge Graphs (SKG), the definitions of « Routes » and « Maps », which remained fuzzy, are progressively clarified (Cabanac et al., 2020) as information retrieval is applied, combined with NLP techniques, to bibliometric information, both to the body of scientific literature (articles, data), but also to the separate information layer of preferences and interpretations expressed on these corpuses (citation, comment, analysis, classification…). The body of scientific literature itself is based on the content of mapped bibliometric knowledge stored in big databases such as the Web of Science, Scopus, PubMed…

Maps have to be organized as being "searchable" (Fabre, 2019) and are efficiently supported by instruments of guidance to help traveling inside their knowledge territories: VOSViewer, for instance, has a long history of providing sophisticated classification features on any article and citation characteristics at any scale, even very large, with the support of a freely available computer program, in an easy-to-interpret way (van Eck & Waltman, 2009).

2. Search Routes: identified communities of multimodal knowledge acquisition

From maps, we observe that researchers draw their own selection within scholarly content through their personal choices of items, all of which belong to a scientific thematic community: in these search "routes", the semantic upgrading of scholarly content can be found and recorded by data on documentary selections, citations, comments, positioning in controversies…

"Routes" within scholarly contents can be clustered in distinct "communities", identified by their own vocabulary, and by their evolution. Routes thus produce data on how researchers exploit and interpret search results, as recorded from their search sessions: data structure of routes on SKGs must record the specific vocabulary features of an identified "community of search" in the incredibly rich and complex network of search data. Section 2 proposes a model clustering communities of search, identified by the relative weight and frequency of scholarly content selections, positioning the selection of communities and their relative dynamics. The results of a test on a large sample of search logs are presented in Section 3.

« Routes » thus trace how communities of users « travel» inside maps, leaving data about their navigation in search sessions and Open Annotations; routes are thus "journeys" marked by additions of local

choices of scholarly contents integrated in SKG, according to preferences in reference chasing, which is based on the personal experience of users (Carevic et al., 2017). Conversational search is recognized as a new way to support research "by focusing on realistic information needs and conversational interactions" (Balog et al., 2001); meanwhile, log analysis of user interactions on shared articles for content recommender systems has been experienced and recorded on a large scale (Moreira et al., 2017).

Upgrading maps according to user preferences recorded from SKG uses, is a narrow path[4] (Zitt et al., 2019). « User based and cognitive approaches to knowledge organization », though well discussed (Hjørland, 2013), suffer from contradictory assessment on their effective impact on the semantic upgrading of research data: a systematic recording of user data in their search sessions represented on a SKG is a field still barely covered today (Fabre, 2019).

3. Routes for resilient semantic upgrade: current innovative open standards

Since 2017 SKG has benefited from the adoption of standards at W3C[5] for an annotation model which observes: « Annotating, the act of creating associations between distinct pieces of information, is a pervasive activity online in many guises ». The STM Publishers Annual Report 2018, mentioned above, notes[6]: « Open annotation shares some features with simpler forms of annotation (e.g., social bookmarking services) but supports multiple annotation types, including bookmarking, highlighting, tagging and commenting. Annotation does not require either the permission from the content annotated website or that it installations of any new software on its part. "

This rich set of open functionalities, free of any predefined link with any business model, allows free and open sharing of knowledge from linked annotations, with services such as the one supplied by OpenEdition.org[7], to which our work refers in the test section of our toolkit (see section 3), or in services like Diigo and PubPeer[8].
Open annotation thus appears among the promising areas for the development of SKG linking documents and preferences. Our testing in section 3 applies such ideas.

4. Routes and the need for new standards of community detection

Tsatsaronis (2020) observes that « a plethora of heterogeneous data sources » interfere in impact assessment of science; the expression of these heterogeneous preferences structured on Routes, does not interfere with scientific publication, as integrated in the Maps (Maisonobe et al., 2018): preferences expressed on Routes allow to record *the ways by which science includes additional pieces of content from human practices of choice and comments.* The related data on preferences record choices and paths between documents on means of discovery: to give an example in other domains, it functions like the ways described for climbing a mountain, while selecting new routes on a wall described elsewhere by geographical maps from aspects of its global structure.

"Human practices" of route-making are themselves a quite heterogeneous category and data interpretation could appear challenging (Zingg et al., 2020), with "trading zones" of multidisciplinary citation exchanges (Grauwin et al., 2012).

[4] Literature identifies 3 ways « ready-made classifications of science, classical information-retrieval searches, mapping and clustering ».
[5] https://www.w3.org/TR/annotation-model/
[6] https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf (p. 171)
[7] https://www.openedition.org/6438
[8] https://www.diigo.com/; https://pubpeer.com/static/about

Attempts to create « route building activities » fuel heated discussions in broad scientific areas, which benefit unevenly from the influences of expertise and research; in molecular biology aera, for instance, with a strong potential for structured services[9], intense discussions are underway on routes of publications (Kahsay et al., 2020), on data routes (Ferrarotti et al., 2019), on comparing routes (Griss et al., 2020).

The fact is that interactions are getting richer every month in this field and that the architecture of information about the routes inside the SKG is part of the necessary knowledge building solution, which interacts with the scholarly contents in a growing "replication crisis"[10]. There is therefore an urgent need for global standards and networks of community standards, connecting the dots of "multimodal knowledge acquisition".

Section 2 SKG GRAPHYP MODELING: DESIGNING SEARCH COMMUNITIES AND THEIR ROUTES

1. Main priorities: interpretability of community searches

A review of the current literature (Chen, Jia & Xiang, 2020) observes: "Recently, reasoning over knowledge graphs has become a hot research topic, since it can obtain new knowledge and conclusions from existing data. » The question then is what kinds of new knowledge are produced in this new context: results could be achieved on new paths of knowledge itself, on the paths of their extraction, and on the interactions of both sides.

Potential ability of graphs to « the common objective to maximizing the knowledge» (Hogan et al., 2020), is clearly entering a new step of research questions « beyond pairwise interactions » (Battiston et al., 2020): newly required approaches will deal with hypergraph structures and « models proposed to generate synthetic structures, such as random and growing bi-partite graphs, hypergraphs and simplicial complexes » that would contribute to a final goal to get « inference from data ». "Inference" could be optimized in recommendation accuracy and also in « interpretability of the recommendation system ». As we underline in our introductory remarks, the recommender system can now offer a "knowledge mapping which enables users to understand the value of recommendations" (Shao et al., 2021).
In this perspective, the goal of our GRAPHYP modeling is to offer the complete vision of all the answers to a keyword to let the user interpret which answer is best suited to his question. We achieve this by providing each user with all "categorisable" answers while showing the user which category their preferences belong to. The latest review of recommender systems (Shao et al., 2021) observes that in this kind of approach, still scarce, which "combines two kind of data into a unified graph to learn the potential characteristics of users and items, and then learn from them", "the feature recommends the appropriate item to the user and gives a reason for the recommendation".

2. GRAPHYP's program: " position your search among others"

The graph structure that we use, hereafter called « GRAPHYP», is a « crown graph » with connected edges[11]. We notice that this hypergraph, as remarked in a recent review (Ouvrard, 2020) reveals a characterization of distance-transitivity[12] that we identify and exploit. A first description of the GRAPHYP modeling can be found in (Fabre, 2019), and in connected papers[13].

[9] https://www.embl.de/research/interdisciplinary_research/bioinformatics/research/index.html
[10] https://www.nature.com/news/replication-studies-bad-copy-1.10634
[11] https://mathworld.wolfram.com/UtilityGraph.html
[12] https://mathworld.wolfram.com/Distance-TransitiveGraph.html
[13] https://www.connectedpapers.com/main/5a00ab293237c4038b9e902adb3fce11ca9e801d/A-Searchable-Space-with-Routes-for-Querying-Scientific-Information/graph

SKG GRAPHYP is built to represent the "searchable space" of a group of search sessions. *It positions a "route of search" from search sessions of a user, and localizes its navigation among possible neighboring routes of other users and items linked to the same search goals*[14].

For a given keyword, assistance to navigation is proposed with detection of knowledge alternative preferences, depending on the weight and frequency of any recorded selection, and marking neighboring search routes showing similarities and oppositions to the selected choice.

While linking these facets of the queried choice, GRAPHYP modeling formalizes a "searchable range" of alternative preferences that can be selected from the same search in identified "communities" of users. With an original positioning of the user's current selection in its whole context, this modeling thus makes it possible to detect "routes" formerly unknown to the user, and allows to compare those routes.

GRAPHYP thus provides a service linking data (articles for instance) and its attributes (readings for instance) and represents communities of search recorded at the global scale of a keyword. Such a service is not yet open to consultation or barely available (Fabre, 2019). With such limitations on alternative documentary selections, « it remains difficult to identify roots of controversies: accurate interpretation of views, downloads, citations, is uncertain » while « searching can sometimes appear as a lonely walk in a forest of hazy homonyms. » (Fabre, 2019).

a. GRAPHYP: community detection on a Bipartite Crown Hypergraph

Preferences of users when reading scientific content will be used here to describe the matrix of possible choices of users. These choices can be recorded between two extreme preferences within the framework of GRAPHYP. Between these extrema, the interaction of a set of typical search sessions helps modeling the complete range of information retrieval needs of distinct communities of users of similarly formulated search sessions. GRAPHYP models all possible choices made from the same question. Our graph building borrowed the unfrequented methodology of a « map equation » formulating a data flow network (Rosvall et al., 2010) described at first in (Rosvall & Bergstrom, 2008), which infers that programmed « links in a network induce movement across the network and result in system-wide interdependence ».

- Recording Search Routes of communities

Let us recall here the methodology proposed in (Fabre 2019) which differs in terms of the calculation bases (see later in the same section).
In order to record « search routes » for a given keyword, which may differ and are meant to be compared, we write:

$$Qn = f(Nn;Kn)$$

where Qn is a number of searches on a given topic. Let us also consider that for each search Q, we can identify a number of users N and a number of items (URL, documents, articles) K recording the search results from a list of items corresponding to keyword. We can consider these users as a community of

[14] « Search goals » as a generic term, encompasses similar queries, keywords, group of URL.

users of the same query route Q[15]. Positioning any distinct route can be expressed for a given search within the limits of a system of typical search sessions.

Let us set two limits of substitution between users and items for any search located within the opposite limit communities: we will consider that one « limit community » is one in which a few users request a lot of items, and the other limit community is one where many users consult a very small number of items.
We can write these possible substitution limits between users and items, N and K:

$$Q1 = f(Nmax, Kmin) \quad or \quad Q2 = f(Nmin, Kmax)$$

Or, in a general form:

$$Q_n = \begin{matrix}[\max, & \min \\ \mathrm{fN}, & K \\ [\max, & \min\end{matrix}$$

We observe that the search function Q described above, coupling N users and K items, can be described in the same way, if for any query Q, one can record pairs of variables expressing different types of search preferences and choices. More generally the search can couple any selection of items N and items K which takes place within the perimeter of a search.

- Recording the dynamics of Search sessions

After testing, we propose here a new method of recording the dynamics of GRAPHYP which differs fundamentally from (Fabre, 2019). Let us calculate the mean value of N and K on the whole set of search sessions: we have the opportunity to normalize the presentation of all search sessions as located above and below the ratio N/K. Let us consider that an additional information on that ratio is given by its dynamics at the scale of the whole set of analyzed search sessions. In fact, with any recorded value of the N/K ratio, there is an associated index of dynamics which expresses that N/K preferences are recorded from an abruptly changing behavior or a steadily increasing or decreasing behavior in reading articles (in our example).

If we decide to represent N/K choices with this additional element of stability/instability of the function, we change our function Q of two variables N, K, into a triplet where the third term linking N and K represents the expression of stability/instability of behaviors of recorded readings of articles. For instance, we could practice community detection of readers of a usual group of articles in chemistry “before” and “after” publication of a new important article and would thus search whether this additional publication accelerate or not readings in peripheral linked domains.

Let us consider that we can measure the value of that third term by a ratio calculated from a value of normalization, which will be expressed from the mean value of N/K.

[15] Alternatively, we could represent Q search session results not by a user/item approach, but by another expression of preferences which could be item/item, where N items and K items are mixable in communities of preferences where we consider that this mix of publications characterizes comparable sets of search sessions.

From that value of normalization of the above-mentioned ratio, let us create a fraction α/β where α is numerator of a fraction calculated from N mean value when reported to β which would be denominator of that fraction from K mean value. This fraction α/β would vary, of course, with any recorded group of values of readers and articles.[16] Note that the value of this fraction can bring a specific element of dynamic analysis, for positioning the respective opposite values of this fraction when observed value tends to one, as normalized on the mean value of the fraction, or when this fraction exceeds its mean value and tends towards infinity.
Fraction α/β makes it possible to measure the "stability" or « instability » of the content either of a search session, when increase or decrease of the quantities N and K are recorded between two searches in the same « route » of searches, or between all routes of search, when N and K values are calculated on the whole set of a group of searches for the purpose of finding communities.

Data structure of any search can now be expressed by an observable and recordable index of change in the relation between the number of users and the number of items. We can thus write:

$$Q_n = f\ [\mathrm{nmax}, \min\ \mathrm{kmax}, \min$$

Let us note that the values « min » or « max" of *N* and *K* provide data on the measured quantity of these variables to « produce" Q, but let us also note that, as observed in (Fabre, 2019) "from one unit of search Q to the other (from one query to the other to the other, for instance), the coefficient of increase or of decrease between value « min" or « max" of *N* and *K* provides an additional index of assessment on the quantity of these variables in the "production" of a search Q." Thus, this additional data provides "a strategic index of characterization of the dynamics of user's preferences: according to « stability" or « instability" of the increases or decreases of those observed indexes of unit variations N and K, and of their ratio, we obtain an original approach of the dynamic behavior of a search community".
The fraction α/β provides a dynamic index of the variations recorded in the practices and controversies of scientific communities, revealing quantitatively how "strong" or "weak" they could be, as measured by the variation of the recorded flow of documentary practices, sudden or steady, strong or weak: by this approach towards sensitiveness to frequency, it could thus help detecting differences between communities approaching the same concept by homonyms[17]. For any unit *n* and *k* in production of Q, the variation of quantities measured by α and β could provide a critical information about Q.

- Networking search sessions

We know that, with the added mix of user and items that it measures, the slope of a set of searches tends to be stable when (α/β) is inferior or equal to 1 or (α / β) tends to 1: as this fraction is tending to its mean value on the whole set of recorded search sessions. The method is here related to the Graph assortativity approaches described in *Wolfram Assortativity*.[18]

[16] Another procedure would be to note α the coefficient of increase of N and β the coefficient of increase of K when Q varies by one unit, that is to say, when an additional query on the same search is recorded. It thus opens an alternate way of calculating the third term.

[17] For the same new category of items, several communities could be « neutral » to a change in publishing orientation, and others could be reactive.

[18] https://reference.wolfram.com/language/ref/GraphAssortativity.html

Conversely, there is also an alternative situation in which (α/β) tends towards + or - infinity, indicating an unstable variation of (α/β), which will increase or decrease sharply: the relations between users and items differ in that case from the average measured by the mean ratio (α/β). So this could indicate that the user and item pair are changing significantly and that a threshold has been reached in the effectiveness of the search process.

In this case, strategically, why continue to "allocate" (n; k) to Q if (α/β) tends toward infinity? This would mean that the combination (n; k) becomes « outperforming", which is questionable when the time comes to decide on a new search Qn+1 with the same characteristics of N and K. With such characteristics for a search, it could, for instance, be more interesting to change N or K than to reproduce the same quantities in the next search. More generally, for transition between Qn-1 and Qn, the choice of stability in N and K could be « justified" if we observe that stability prevails with (α/β) tending to 1 and « questionable" when instability prevails with α/β tending to + or- infinity.

Over the course of a long series of search, there will be a learning process that will give meaning to the expression of limits in the variations of the proposed investigation. From this perspective, it is possible to set arbitrary limits of variations for a given panel of search sessions, with the « best" and the « worst", for instance:

$$\text{Best} = N^{\alpha}_{\max}, (\alpha/\beta) \to 1, K^{\beta}_{\min}$$

And

$$\text{Worst} = N^{\alpha}_{\min}, (\alpha/\beta) \pm \infty, K^{\beta}_{\max}$$

If we have selected these two opposite limits to search sessions belonging to an STI data search, we need to give details on the calculation of our limits.

How to select the minimum and maximum values of N and K? How to identify "best" and "worst" cases?

There are several steps to prepare for the calculation of these values:

- SELECTION OF PARAMETERS FOR N AND K:

these values in (Fabre, 2019) correspond to the number of users and the number of articles, but it could be any pair of data that can be combined to express preferences. For instance: number of articles citing the article of Author X and number of articles citing author Y; or, number of readers of article A and number of readers of article B, etc.

- SELECTION OF MIN AND MAX VALUES OF N AND K:

when data on N and K are recorded, then one has to identify their respective minimum and maximum recorded values. The four values N (max & min) and K (max & min) will then give the quantitative limits of GRAPHYP recording, which supposes giving the location of all intermediate values between the two limits recorded for a search session. Min and max value of N and K are findable and the return of GRAPHYP is better when recorded from rather large data series.

- SELECTION OF MEAN VALUE OF COEFFICIENTS α AND β

The mean fraction is calculated on the whole group of searches: it can provide a value for "positioning" the respective opposite values of this fraction at the scale of the whole group of searches, in association with mean and max values of N and K, and, from this basis, on the whole panel of search sessions between the min and max values of N and K.

- SELECTION OF « BEST » AND « WORST » CASES

One has to select, from the user's point of view, GRAPHYP results that one considers as « objective » limits of the analysis. Those would be the « best » and « worst » situations between which one expects to know how all users of a search session are positioned. Notice that search sessions can be recorded over time for an individual user, or for a comparative set of communities of users.

When one expects assessing preferences of users of a search session, it is easy to observe if, for instance, Nmin is better than Kmin as the « best » case. In addition, the min and max preferences tested on limits can be modified and the results recorded by GRAPHYP can be tested on other limits, provided that the preferences are formulated with opposite symmetrical choices in order to let GRAPHYP logics work, which supposes having limits for the classification of recorded results.

One can then write that:

"Best"	"Worst"
N_{max} $(\alpha/\beta) \rightarrow 1$ β K_{min}	N_{min} $(\alpha/\beta) \pm \infty$ β K_{max}

- Positioning search sessions on routes: a "GRAPH AS A COMPASS"

Based on the above-mentioned expression, our two triplets give shape to a formal graph-based representation of paths existing between the two limits and possessing the characteristic of positioning all the non-contradictory solutions existing between these two limits, included within three parameters. We selected the following bipartite crown graph with connected nodes, which is the unique way to give the needed representation by a typical network.

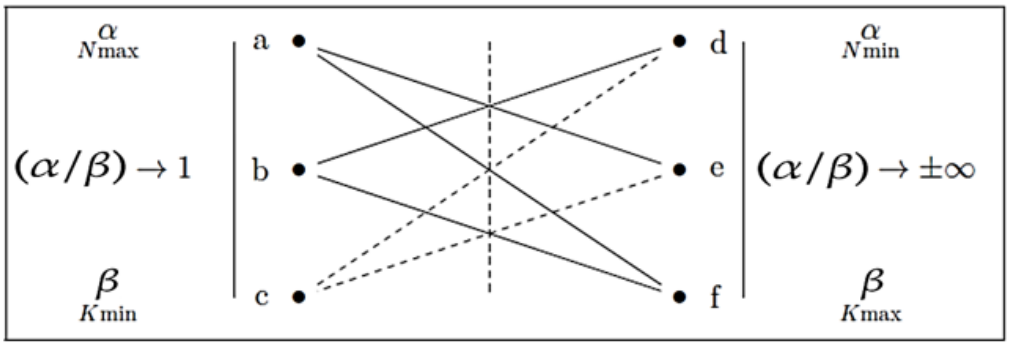

Figure 1. "Compass" GRAPHYP: Positioning Typical Search Sessions

Adapted from (Fabre, 2019)

Figure 1 shows a complete representation of all the typical intermediary situations between our two limits, which gives us a tool for the classification of observed search sessions Q in series of searches on a given keyword, according to the user and item choices as tending to above fixed limits. Structured by GRAPHYP, this set of search session typical positions has two main characteristics, which will be detailed below with the help of Table 1.

As observed in Table 1, GRAPHYP expresses the whole set of *non-contradictory* positions that structured search routes could occupy during the analysis of a group of searches on a given keyword, for instance, on reading articles.

NODE (Query)	TYPICAL DIRECTION OF SEARCH (Users, Items)
a	$N\alpha_{\max}, K\beta_{\max}, (\alpha/\beta) \to \pm\infty$
b	$N\alpha_{\min}, K\beta_{\max}, (\alpha/\beta) \to 1$
c	$N\alpha_{\min}, K\beta_{\min}, (\alpha/\beta) \to \pm\infty$
d	$N\alpha_{\min}, K\beta_{\min}, (\alpha/\beta) \to 1$
e	$N\alpha_{\max}, K\beta_{\min}, (\alpha/\beta) \to \pm\infty,$
f	$N\alpha_{\max}, K\beta_{\max}, (\alpha/\beta) \to 1$

Table 1. Classification of Typical Search Sessions in GRAPHYP

Adapted from (Fabre, 2019)

For instance, from Figure 1 and Table 1, node c is linking hedge cd and hedge ce: this node is creating the analytic position listed in Table 1, with all typical positions that can be recorded in GRAPHYP data structure, between a and f positions of search sessions Q: they find their location in this graph as tending to min max values of N,K.

When recording data on a set of search sessions, and preparing them for GRAPHYP uses, we can apply the classifications proposed above to represent, for users, all of the practices recorded during search sessions issued from the same scientific theme. In section 3, we test the conditions to process data on readings in the context of scientific impact assessment.

The following example shows all the logic at work on a reduced panel of 100 search sessions; we could accurately account for the practices of article readers during that recording of search sessions, using GRAPHYP:

- Class each of that 100 Q search sessions between our limits as grouping together a number of readers N of articles K and expressing a ratio of readers to articles that may be above or below the calculated median value of users and items, during the recording of these search sessions (tending to max or to min value of N and K).
- Asses the stability/instability of recorded practices: Separate our reading practices between those which, for the expression of their values, expressed a behavior differing of mean value and presenting stability or change in reading practices (for instance before and after publication of an important article) by recorded value of the fraction α/β, referred to mean value of that fraction.
- Position the incoming practice of search session of a new reader of articles, and give that reader the whole landscape of all others reading practices in this community.

GRAPHYP thus helps transforming KG from its function of classification into a tool for positioning present human choices with the aim of future choices, as a kind of a « compass » for navigating in a « searchable » space (Fabre, 2019). At least GRAPHYP makes it possible to find the closest triplet in a set of vectors, while choosing a route in a multidimensional space. GRAPHYP therefore belongs, even to a large extent, to community detection (see further, functionalities) approaches (Waltman & van Eck, 2013). Community detection with efficient identification of cliques inside a hypergraph, can be represented in GRAPHYP as follows:

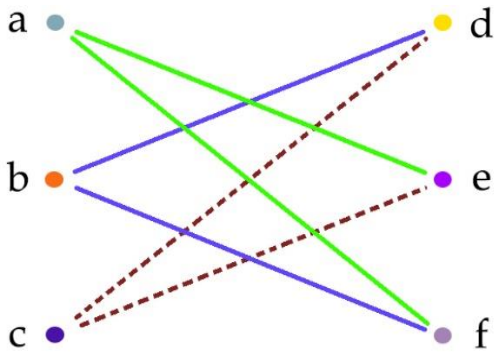

Figure 2. Markings in GRAPHYP

- On any node are recorded the numbers of Q search sessions that appear to belong to that type
- Any « type » of search session can be represented by separate color for any node, (for instance, color intensity could vary with the number of Q sessions loaded for that node)
- Most dense type (here a) is designed with hedges on a continuously colored design (here in green because we could note that is the color of search sessions of the **a** type)
- Newly incoming search is represented by a discontinuous line (here on node c) to give characterization of the current search session: it allows the incoming user to get three elements of information: 1 is the category to which the search belongs, 2 the whole representation of all search sessions, which allows to compare his practice with all the others, 3 the "dominant" profile of group of search sessions (here in green with "a node" search sessions).

With the help of functional tools (see point 3 hereafter) the user can see all the "routes" of search sessions recorded in GRAPHYP, and navigate inside, just like with a navigation compass, and its own route can be traced.

In Section 3, we will present an analysis of the results, tested from our modeling and GRAPHYP representation, for the data of about 1 million searches. It can be noted that "click metrics analysis" of scholarly contents begins to develop (Fang et al., 2021).

3. GRAPHYP structure: coding a bipartite graph with data

The mutation from two variables into triplets is modeled to allow GRAPHYP to manage data flows corresponding to its structure as a crown hypergraph.

This section deals with the Hungarian method or the Hungarian algorithm. More precisely, it can be assigned to linear optimization. The Hungarian algorithm is a method for solving unweighted and weighted assignment problems in bipartite graphs. We also deal with weighted matchings. The problem of finding a matching with maximum weight in a bipartite graph is equivalent to the assignment game. The specific method of creating graphs for hypergraphs described in this article is applied to the GRAHYP structure with an example with data. **A bipartite graph with a Python code has been developed and is available on Github under the following link**[19].

We have developed this code with a simple example and it can of course be used for other data, but only the "Domain" and "Subdomain" have to be supplemented with the desired data. Various problems in computer science can be solved with the help of graphics. For example, the problem of the formation of object pairs in a graph, in which the objects that can possibly be paired are connected by edges

[19] https://github.com/OtmaneAzeroualDZHW/Creating_bipartite_graph

(Schubert, 2012). These edges mostly represent settings that can collide with one another. There are applications in particular for this problem (e.g., bipartite matching and Hungarian method).

Figure 3 shows the result of the code.

```
('A', 'F', 'weight':  10)
('B', 'F', 'weight':  12)
 ('A', 'E', 'weight':  4)
('C', 'E', 'weight':  67)
('B', 'D', 'weight':  27)
('C', 'D', 'weight':  33)
```

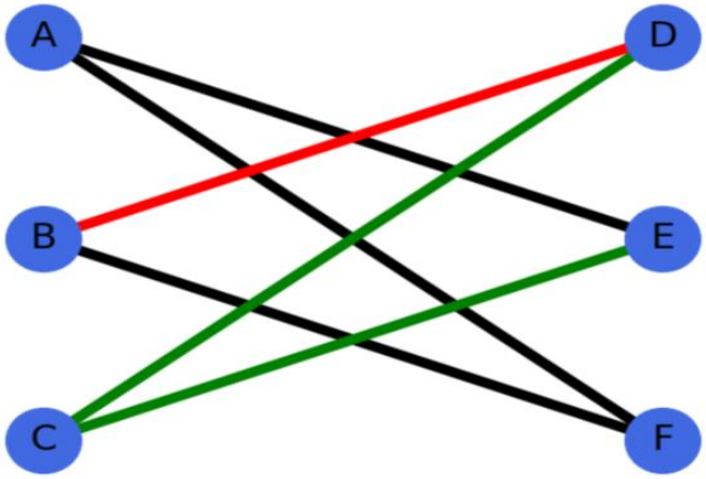

Figure 3. Bipartite Graph with Hungarian method

In this example an equality graph is generated where each edge weight is equal to the sum of the marks of both neighboring nodes. In order to determine the equality graph, the algorithm has to find the maximum neighboring edge of each node and assign its weight to the node as a marker. The equality graph is determined based on the node markings (black). In order to carry out the matching, the algorithm must first determine an augmentation path. The matching could be supplemented using the augmentation path found. This can be seen in the green and red between the elements.

GRAPHYP could allow users to have a picture of the reading practices of an identified community of readers while recording and visualizing their practices. Those functionalities of GRAPHYP are described in next section.

4. EXPLORING ROUTES OF SEARCH WITH GRAPHYP

Function 1: POSITIONING A SEARCH ROUTE

As a sort of "digital compass" GRAPHYP provides all possible "directions of search" and locates the recorded direction of an individual search. It thus implements an original approach to "path-based reasoning" (Jagvaral et al., 2020).

This set of possible "positions" during a search gives an overview of the proposed modeling of the searchable space for STI queries, according to six typical search routes. Figure 4 shows how GRAPHYP allows to explore typical Neighbor Search Sessions, in a context of « betweenness centrality »20.

[20] https://medium.com/rapids-ai/rapids-cugraph-networkx-compatibility-d119e417557c

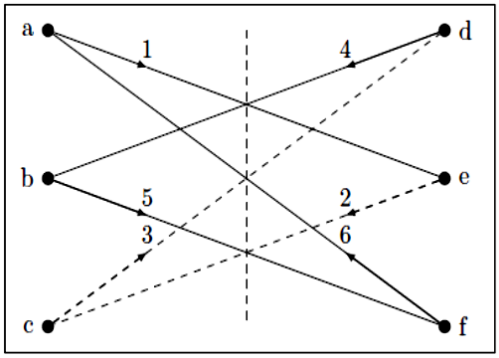

Figure 4. Exploring Neighbor Search Sessions with GRAPHYP "Compass"
Source: (Fabre, 2019)

Search sessions can thus be interconnected and explored, as a continuous circulation from node to node showing a common edge in all pairs of edges corresponding to any node. A circular « travel » is possible inside GRAPHYP, from edge to edge, while coming back to the starting node by the way of the complementary edge to that node.

Bipartite graph analysis is an alternative to reveal clusterization in complex systems (Palchykov & Holovatch, 2018), as exploring and analyzing sessions of users could benefit from results produced from the GRAPHYP data structure. We note that research on user-items connections has not currently followed this path and is mostly bound to classical approaches to recommendation applied to SKG (Wang et al., 2018). Recently Graph Convolutional Networks (GCN) "and variants have achieved state-of-the-art results on classification tasks, especially in semi-supervised learning scenarios. A central challenge in semi-supervised classification consists in how to exploit the maximum of useful information encoded in the unlabeled data » (Pedronette & Latecki, 2021).

The original structure of GRAPHYP thus proposes an *undirected* weighted graph structure (each edge is bi-directional and receives a distinct weight from the nodes linked inside a network), which is connected (one can reach any node from all other nodes inside identified paths) and builds a minimum spanning tree (MST) as a subgraph containing all nodes, connected here with the minimum possible number of edges. GRAPHYP gives then a structural approach to least distances solutions in bi-partite graphs representations (it could be borrowed by analyses developed in applications of convolutional networks).

Function 2: MAPPING SEARCH ROUTES

As described in (Fabre, 2019) users of GRAPHYP will be able to learn from their past recorded behavior as well as from viewing the recorded routes of other users of the same base, by identifying their "search position" and their "search route": these goals could be reached from a grid of structured data on the recorded searches of all users, as shown in Figure 5.

SKG GRAPHYP could thus be used as a "compass", recording and directing digital navigation in an environment of articles and other scientific resources.
As explained above, the pairs of edges in GRAPHYP always belong to a given node (here a, b, c, d….). **Any node designs a grid of proximities between partially neighboring nodes**: we will then exploit those partial proximities to identify paths which, by using next neighbor circulation paths, can:

- allow identifying any optimal "route of preferences" of a user, according to relevant identified proximities of edges and nodes;

- record any observed path of users which, during their past search sessions, have used a recorded way.

Figure 5 shows a genealogy that could be recorded from data on search routes, analyzing paths of discovery by assessment of scientific documents characterized by community detection, in labs or teams for instance. Any of the distance evaluation methods can be applied to this architecture[21] in order to explore neighboring routes, their genealogy, and mapping distances and differences between search sessions.

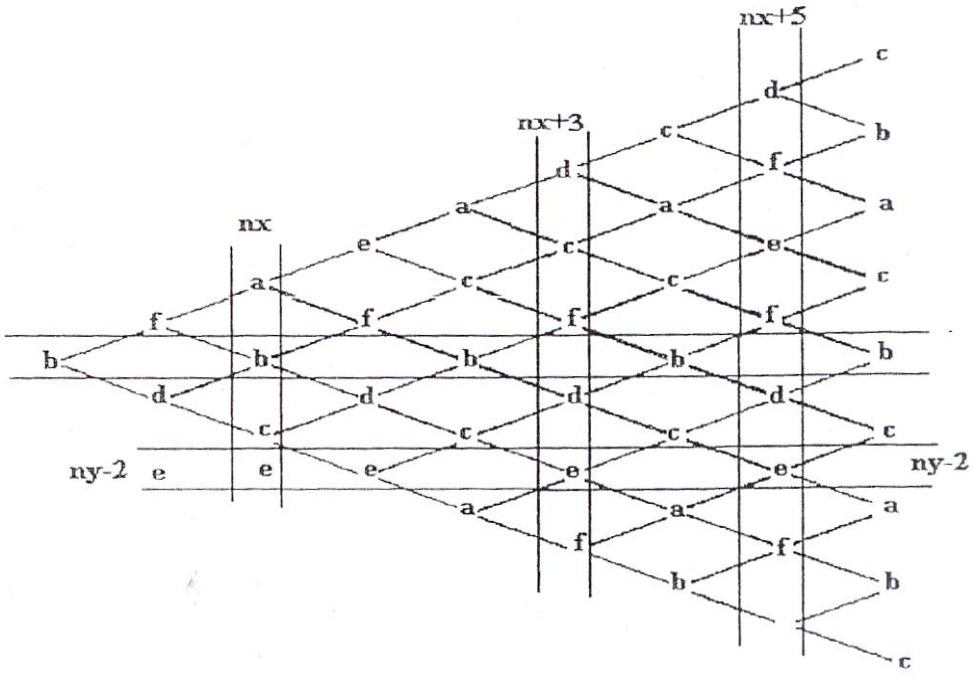

Figure 5. Positioning recorded nodes on chronologies of data search routes

Source: Fabre (2019)

In SKG logics, it could be applied to the comparison of search routes to discovery: documentary choices (readings, downloads, comments…) could provide data to study the appearance of neologisms and changes in vocabularies. This type of a grid could also allow projections from a given node, on the positioned nodes of closest neighbor and thus provide a basis for a future search strategy.

Characteristic 3: SCALING KNOWLEDGE EXTRACTION: SELF SIMILAR GRAPHYP

The mapping of STI "searchable routes" is thus designed with rules equivalent to air or sea routes. Cooperative or connecting routes can be identified by the GRAPHYP data structure. A final property of self-similarity, shown in Figure 6 below, can help tracing routes by providing a tool for scaling and zooming on GRAPHYP search sessions, selected at the relevant scale.

GRAPHYP's geometry can be duplicated at any scale: as illustrated by Figure 6 below, GRAPHYP with nodes A, B, C, etc. is built by addition of graphs of the same shape: this self-similarity characteristic of GRAPHYP (Fabre, 2019) allows knowledge extraction at any scale and allows operating scalability from perimeters of information processed from operations of addition, subtraction, multiplication and division.

[21] https://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures.html

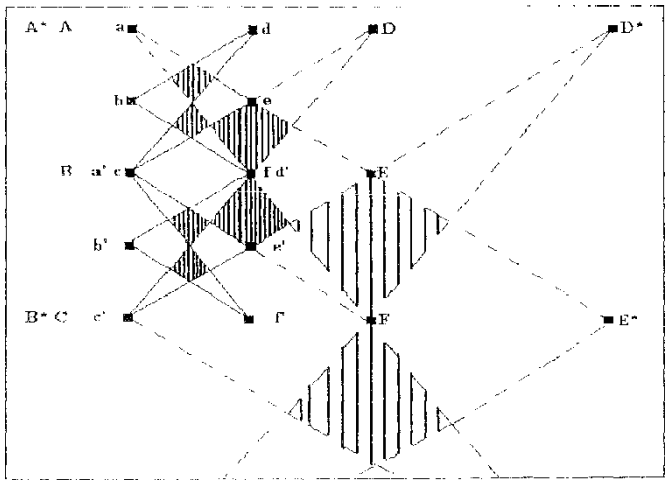

Figure 6. Searching at Various Query Scales: Fractal Scalability of GRAPHYP

Scalability of GRAPHYP provides therefore a tool connecting approaches identified in different layers of the same object of information; in SKGs applications it could be particularly helpful to zoom inside scientific field stages or even to compare evolutions of a given term used by different scientific fields. To the best of our knowledge, this property of self-similarity is not yet highlighted in current reviews of research on design of KGs (Ji et al., 2020)[22]. Such an architecture of self-similar graph could thus mobilize in an original way, the methods described in Wolfram for representation of Graph communities, like Random Graph or Community GraphPlot[23] . It could also give an additional way of solution to problems of Distributed Large-scale Natural Graph Factorization (Ahmed et al., 2013) and even supply a framework for large scale graph decomposition and inference and give a way of research on "automating the expansion of knowledge graphs" (Yoo & Jeong, 2020) in directions where bifurcations in sense building have to be found, and thus mainly in SKGs applications.

Section 3 TESTING GRAPHYP: COMMUNITY DETECTION IN READINGS OF SCHOLARLY CONTENTS

1. The need for "Open Process" principles for design of search preferences

Services for assistance to navigation on SKG are unevenly developed. Assistance exists in library management[24], with a wide range of services and search techniques in contexts for "solving the problem of problem solving"[25] and on a range of services for management of logistic of documentation which re-use parsing data, like Ezparse[26]. Main purpose in those cases remains library management services and contract management with publishers delivering contents from their data bases, and recording uses in their own "publisher's knowledge books".

There is no clear delineation on how these current developments could directly benefit the choices of researchers: they could not benefit as users of their own data of search session, to orient their work with uses of representations of their routes by SKG.

However, Industry is already long aware of community detection in analysis of scientific contents and applies it to development of services to users, exploiting a wide range of features: readings are one of them.

[22] http://shaoxiong.ml/knowledge-graphs/

[23] https://reference.wolfram.com/language/ref/FindGraphCommunities.html

[24] https://www.ebsco.com/products/ebsco-usage-consolidation

[25] https://www.lens.org/

[26] https://www.ezpaarse.org/

" Article Recommender in ScienceDirect suggests articles that are relevant to the individual user based on what they're reading. To make these recommendations, data from millions of other researchers who've read the same article is anonymized, aggregated and analyzed in milliseconds, resulting in suggestions that save that person a significant amount of time» ; altogether, service is also available for usage data: « With usage data, it could even be possible for researchers to learn from what their networks have downloaded or saved. If done well – respecting privacy and confidentiality – this would help researchers find relevant information even more quickly. ».[27]

2. Our testing

In order to identify the nature of the data required to implement Graphyp in the context of a digital library, we made a first prototype using access log files from OpenEdition.org platforms[28]. These log files were collected by the web analytics platform Matomo[29] and then filtered in order to eliminate connections from bots and requests for files other than papers. User IDs have been associated to the requests to the Web server according to the anonymized IP address and sessions IDs have been added based on the recorded timestamps.

This means that in the following we call session a sequence of actions (content requests) performed by an anonymous user in a limited time. We can assume that each session thus corresponds to a specific need, even if we do not know the original request the reader made (the referers haven't been communicating queries for a few years now and the incoming links are usually very vague). In the vast majority of cases, the reader comes directly from a Web search engine and then follows the links in the pages corresponding to the retrieved papers.

We thus have for each session a time-ordered sequence of articles read. Since the queries here are unknown, we can only have the ambition to compare general reader behaviors, regardless of their information needs.

The process is then as follows.

First of all we divide all the logs into blocks of sessions. Then, for each session, within each block, we estimate the values of K as the number of articles read by a reader.

This makes it possible to estimate, in a given block, the number of readers N who have read K articles and thus the mean values of the different N and K for this block. α and β values as well as their ratio are calculated, for each bloc of sessions except the first one, from the mean values of N and K.

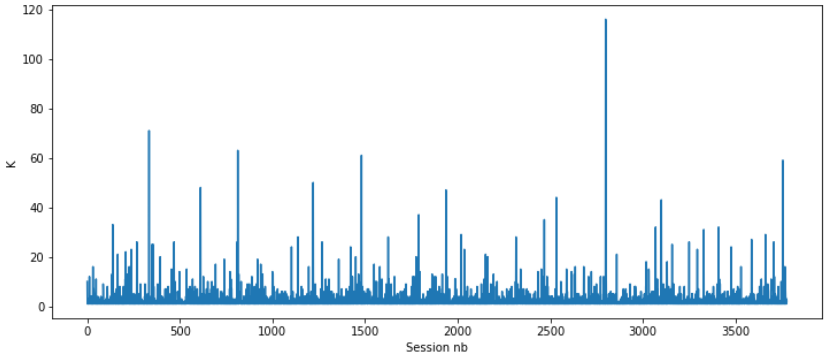

Figure 7. K values for about 4000 sessions

[27] https://www.elsevier.com/connect/respecting-data-privacy?SQ_VARIATION_149028=0 Retrieved 20 Dec 2020
[28] https://www.openedition.org/10918?lang=en
[29] https://matomo.org/about/

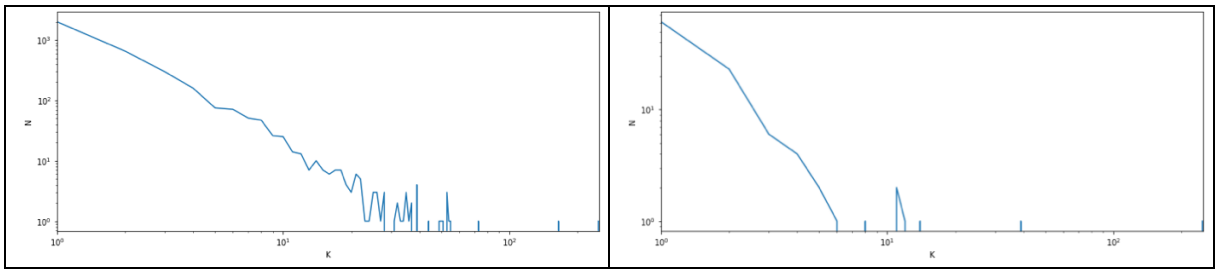

Figure 8. - N and K values for two different blocks of sessions (on the left, quite a few people read more than 10 articles while on the right the vast majority of readers read only one article).

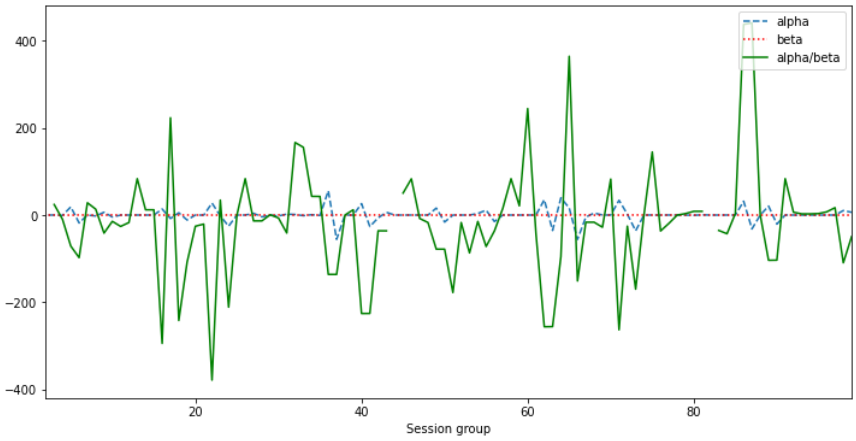

Figure 9. α, β and $\frac{\alpha}{\beta}$ for 100 blocks of 10 000 session

Once this is done, it remains to determine, for each session its type according to its associated values of N, K and α / β.

Some thresholds make it possible to identify the tendencies towards the min/max values of N and K and towards ±infinity or stability (1) for $\frac{\alpha}{\beta}$ (for example, we can choose to consider K as type K_{max} if $K > K_{max} - (K_{max} - K_{min}) \times z$ with $0 < z < 1$).

Lastly, the sequence of the types gives us the routes of search. By considering 100 blocks of 10 000 sessions corresponding to 1 million documents read, 42% of the sessions are type b. Notably because the blocks are not built according to the user queries (which are not known here), this result was expected: the majority behavior is stable.

Section 4 FURTHER WORKS

We plan to develop GRAPHYP with reference to scalability.

Another track would be to clarify the status and meaning of data itself, document, information and knowledge, and their use by the SKG modeling. What is the difference between “embedded knowledge” (= represented knowledge in documents) and generated knowledge? These questions are anything but new (Bates, 2005; Buckland, 1991; Frické, 2019; Furner, 2016), but the importance of the answers is growing, and GRAPHYP modeling should help to build up comparative approaches to practices.

At least, our current tracks on “Search and select data processing” as part of a global modeling of data on annotations interfacing research practices and content updating, should be developed in two main directions:

- Optimization of SKG representations for research impact assessment: results of the search activities should be better identified, shared, and a more robust and clear help for navigation on knowledge maps should be designed.

- Partition: exploration of theoretical and operational links between modeling of research contents and search results has to be clarified, when they interact on the data structure of an SKG.

CONCLUSION

SKGs have been described as large networks of entities and relationships that focus on the scholarly domain and typically contain metadata describing research publications such as authors, venues, organizations, research topics, and citations (Buscaldi et al., 2019). While a knowledge graph "acquires and integrates information into an ontology and applies a reasoner to derive new knowledge" (Ehrlinger & Wöß, 2016), a SKG "represents scientific information" (Auer et al., 2018). Its main purpose is to provide assistance to cope with the overwhelming volume, variety and velocity of research information.

What is in fact the perimeter or SKG, and to what extent does it encompass the uses of science?

At least, we have observed that Knowledge is a path, not a piece of information: it includes delineations and specificities of reference chasing which give an individual imprint to the results. In that way SKG is shifting from processing standardized information to representing debates, with a process of mutual education and training.

In these directions, with Michael Jordan[30], we share the idea that « We will need well-thought-out interactions of humans and computers to solve our most pressing problems. And we will want computers to trigger new levels of human creativity, not replace human creativity (whatever that might mean)".

Our answer in this article is that there is room for data structure combining documents and annotations in modeled communities. Yet we observe that "Search is not Research" and further work should thus explore links between modeling, data recording of research contents, and "search and select" results in the SKG data structure.

Author contributions

Renaud Fabre: Conceptualization, Writing - original draft. Otmane Azeroual: Formal analysis, Writing - original draft. Patrice Bellot: Investigation, Writing – original draft. Joachim Schöpfel: Supervision, Writing - original draft, review & editing. Daniel Egret: Validation, Writing – review & editing.

Competing interests

The authors declare no competing interests.

References

Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., & Smola, A. J. (2013). Distributed large-scale natural graph factorization. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). Association for Computing Machinery, New York, NY, USA, 37–48. DOI: https://doi.org/10.1145/2488388.2488393

[30] https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9

Aryani A., Fenner M., Manghi P., Mannocci A., & Stocker M. (2020). Open Science Graphs Must Interoperate!. In: L. Bellatreche et al. (Eds.), ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. Communications in Computer and Information Science, vol 1260 (pp. 195-206). Springer. https://doi.org/10.1007/978-3-030-55814-7_16

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., & Vidal, M. E. (2018). Towards a Knowledge Graph for Science. Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, 1–6. https://doi.org/10.1145/3227609.3227689

Bates, M. J. (2005). Information and knowledge: An evolutionary framework for information science. Information Research, 10(4), Paper 239. Retrieved from http://www.informationr.net/ir/10-4/paper239.html

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J., & Petri, G. (2020). Networks beyond pairwise interactions: Structure and dynamics. Physics Reports, 874, 1-92. https://doi.org/10.1016/j.physrep.2020.05.004

Buckland, M. K. (1991). Information as thing. Journal of the American Society for Information Science, 42(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Buscaldi, D., Dessì, D., Motta, E., Osborne, F., & Reforgiato Recupero, D. (2019). Mining Scholarly Publications for Scientific Knowledge Graph Construction. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 8–12). https://doi.org/10.1007/978-3-030-32327-1_2

Cabanac, G., Frommholz, I. & Mayr, P. (2020). Scholarly literature mining with information retrieval and natural language processing: Preface. Scientometrics, 125, 2835–2840. https://doi.org/10.1007/s11192-020-03763-4

Carevic, Z., Lusky, M., van Hoek, W., & Mayr, P. (2017). Investigating exploratory search activities based on the stratagem level in digital libraries. Int J Digit Libr 19, 231–251. https://doi.org/10.1007/s00799-017-0226-6

Chen, X., Jia, S. & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications,141, 112948; https://doi.org/10.1016/j.eswa.2019.112948

Chen, J., Yang, L., Raman, K., Bendersky, M., Yeh, J., Zhou, Y., Cai, D., & Emadzadeh, E. (2020). DiPair: Fast and Accurate Distillation for Trillion-Scale Text Matching and Pair Modeling. in "Findings of the Association for Computational Linguistics: EMNLP 2020", Online, pp. 2925-2937. doi = "10.18653/v1/2020.findings-emnlp.264",

Ehrlinger, L., & and Wöß, W. (2016). Towards a definition of knowledge graphs. SEMANTiCS 2016, Posters and Demos, 48, pp. 1–4. http://ceur-ws.org/Vol-1695/paper4.pdf

Fabre, R. (2019). A searchable space with routes for querying scientific information. In Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019), 112-124, http://ceur-ws.org/Vol-2345/paper10.pdf

Fang, Z, Costas R, Tian, W., Wang, X., & Wouters, P. (2021). How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. J Assoc Inf Sci Technol. 1–15. https://doi.org/10.1002/asi.24458

Ferrarotti, M.J., Rocchia, W. & Decherchi, S. (2019). Finding Principal Paths in Data Space. In IEEE Transactions on Neural Networks and Learning Systems, 30(8), 2449-2462. http://doi.org/10.1109/TNNLS.2018.2884792

Feyerabend, P. (1975). Against Method: Outline of an Anarchistic Theory of Knowledge. London, New York: New Left Book.

Frické, M. (2019). Knowledge pyramid. The DIKW hierarchy. ISKO Encyclopedia of Knowledge Organization. Retrieved from https://www.isko.org/cyclo/dikw

Furner, J. (2016). " Data ": The data. In M. Kelly & J. Bielby (Eds.), Information Cultures in The Digital Age: A Festschrift in Honor of Raphael Capurro (pp. 287–306). Wiesbaden: Springer.

Grauwin, S., Beslon, G., Fleury, É., Franceschelli, S., Robardet, C., Rouquier, J.-B. & Jensen, P. (2012), Complex systems science: Dreams of universality, interdisciplinarity reality. J Am Soc Inf Sci Tec, 63, 1327-1338. https://doi.org/10.1002/asi.22644

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., & Hermjakob, H. (2020). ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. Molecular & Cellular Proteomics, 19 (12), 2115-2124. https://doi.org/10.1074/mcp.TIR120.002155

Hjørland, B. (2013). User-based and Cognitive Approaches to Knowledge Organization: A Theoretical Analysis of the Research Literature. Knowledge Organization 40, no. 1: 11-27. Also available in ISKO Encyclopedia of Knowledge Organization, ed. Birger Hjørland, coed. Claudio Gnoli, http://www.isko.org/cyclo/user_based

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G. et al. (2020). Knowledge graphs. arXiv: https://arxiv.org/abs/2003.02320

Jagvaral, B., Lee, W., Roh, J., Kim, M., & Park, Y. (2020). Path-based reasoning approach for knowledge graph completion using CNN-BiLSTM with attention mechanism. Expert Systems with Applications, 142, 112960. https://doi.org/10.1016/j.eswa.2019.112960.

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., & Auer, S. (2019). Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19). Association for Computing Machinery, New York, (pp. 243–246). DOI: https://doi.org/10.1145/3360901.3364435

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2020). A Survey on Knowledge Graphs: Representation, Acquisition and Applications. ArXiv: https://arxiv.org/pdf/2002.00388.pdf

Kahsay, R., Vora, J., Navelkar, R., Mousavi, R., Fochtman, B. C. et al. (2020). GlyGen data model and processing workflow. Bioinformatics, 36, 12, 15 June 2020 (pp. 3941–3943). https://doi.org/10.1093/bioinformatics/btaa238

Kejriwal, M. (2020). Knowledge Graphs and COVID-19: Opportunities, Challenges, and Implementation. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e45650b8

Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago: Chicago University Press.

Liu, Y., Wang, Q., Wang, X., Zhang, F., Geng, L., Wu, J. & Xiao, Z. (2020). Community enhanced graph convolutional networks. Pattern Recognition Letters, 138, 462-468. https://doi.org/10.1016/j.patrec.2020.08.015

Lü, L., Chen, D., Ren, X., Zhang, Q., Zhang, Y., & Zhou, T. (2016). Vital nodes identification in complex networks. Physics Reports, 650, 1-63. https://doi.org/10.1016/j.physrep.2016.06.007.

Maisonobe, M., Jégou, L., & Cabanac, G. (2018). Peripheral forces: The growing impact of second-tier cities is narrowing the gap in research production [Comment]. Nature, 563 (7729), 18-19. https://doi.org/10.1038/d41586-018-07210-6

Moreira, G. (2017). Articles sharing and reading from CI&T DeskDrop.https://www.kaggle.com/gspmoreira/articles-sharing-reading-from-cit-deskdrop

Ouvrard, X. (2020). Hypergraphs: an introduction and review. ArXiv:2002.05014. https://arxiv.org/abs/2002.05014

Palchykov, V. & Holovatch, Y. (2018). Bipartite graph analysis as an alternative to reveal clusterization in complex systems. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv (pp. 84-87). https://doi.org/10.1109/DSMP.2018.8478505

Pedronette, D.C.G., & Latecki, L. .J. (2021). Rank-based self-training for graph convolutional networks. Information Processing & Management, 58(2), 102443. https://doi.org/10.1016/j.ipm.2020.102443

Rosvall, M., Axelsson, D. & Bergstrom, C.T. (2010). The map equation. Eur. Phys. J. Special Topics 178, 13–23.

Rosvall, M. & Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. PNAS 105, 1128-1123. https://doi.org/10.1073/pnas.0706851105

Schubert M. (2012). Paarungsprobleme und ihre ungarischen Lösungen. In: Mathematik für Informatiker. Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-1995-6_19

Shao, B., Li, X., & Bian, G. (2021). A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. Expert Systems with Applications, 165, 113764. https://doi.org/10.1016/j.eswa.2020.113764.

Thelwall, M. (2020). Mendeley reader counts for US computer science conference papers and journal articles. Quantitative Science Studies, 1(1), 347-359. 10.1162/qss_a_00010

Tsatsaronis, G. (2020). Metrics and trends in assessing the scientific impact. In BIR 2020 Workshop on Bibliometric-enhanced Information Retrieval. In press. http://ceur-ws.org/Vol-2591/paper-02.pdf

van Eck, N. J. & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538. http://doi.org/10.1007/s11192-009-0146-3

Waltman, L., & van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B, 86, 471. https://doi.org/10.1140/epjb/e2013-40829-0

Wang, X., Wang, D., Xu, C., He, X., Cao, Y. & Chua, T.S. (2018). Explainable Reasoning over Knowledge Graphs for Recommendation. In The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 5329-5336. ArXiv: https://arxiv.org/pdf/1811.04540.pdf

Yoo, S., & Jeong,O. (2020). Automating the expansion of a knowledge graph. Expert Systems with Applications, 141, 112965, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2019.112965.

Zingg, C., Nanumyan, V., & Schweitzer, F. (2020). Citations driven by social connections? A multi-layer representation of coauthorship networks. Quantitative Science Studies. Advance publication. https://doi.org/10.1162/qss_a_00092

Zitt, M., Lelu, A., Cadot, M., & Cabanac, G. (2019). Bibliometric Delineation of Scientific Fields. In: Glänzel W., Moed H.F., Schmoch U., Thelwall M. (Eds) Springer Handbook of Science and Technology Indicators. Springer. https://doi.org/10.1007/978-3-030-02511-3_2