The Open Web of Science

Making science browsable

Open Scholar CIC & Future of Research

1 Background

The volume of knowledge is growing at an exponential rate, but so is its rate of fragmentation. It has become extremely hard, if not impossible, for individual researchers to keep up with all the relevant literature in their field. Time is increasingly being invested in developing search engines with ever more sophisticated algorithms in an attempt to make sense of the articles' content, and match it with user expectations. Given the human problem of reducing knowledge to a correct set of representative keywords, artificial intelligence cannot yet extract "meaning", assess "quality", or predict "impact", and is not ready to mediate these core values in the process of scientific communication and evaluation. Algorithms cannot reproduce the complexity and efficiency of scientists themselves instinctively searching for relevant papers, while at the same time making the necessary value judgements. In the case of data and software, these problems are compounded and the absence of tools is leading to a loss of visibility, reusability, efficiency and reproducibility. What is needed is **an efficient classification and structuring of global scientific knowledge** utilizing the power of technological innovations in the area of open science and **the collective intelligence** of research scientists themselves.

2 Our proposed solution

We propose crowdsourcing the research community's collective wisdom to build a complete *semantic web* for science with the necessary tools for it to grow into a global portal. We will create a place for scientists to develop a glossary of keywords, with vertical (parent-child type of relations such as Biology → Developmental biology) and horizontal links (interdisciplinary connections such as Protein structure ↔ Nuclear Magnetic Resonance). By self-organizing and self-regulating the evolution of this "**Open Web of Science**" (OWS), scholars will build the everyday consensual vocabulary that will enable them to find what they are looking for. In this logical framework, **all** scientific digital objects, past and future, including articles, datasets, software, blogs, events, presentations and more, will be tagged with the most relevant keywords in order to be accessed *exhaustively and unambiguously* with a single click using an intuitive graphical interface. Our aim is to make scientific knowledge intelligible and navigable, and to reverse the process of fragmentation. Importantly, OWS can become a great educational tool for opening the door to scientific knowledge for the general public.

3 Tools and user experience (UX)

3.1 Browsing the Open Web of Science

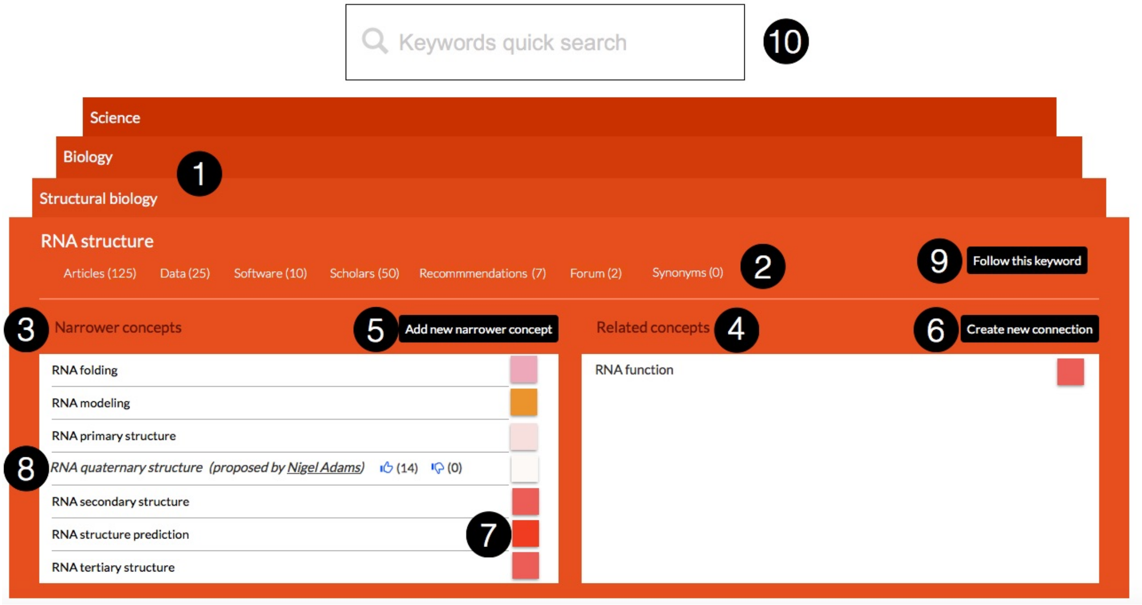


Figure 1: Example interface (design in progress)

Anyone can use the OWS interface to browse and access scientific content.

Starting from the parent node 'Science', users can explore topics of interest by following the trail that leads them into more specialized knowledge (narrower concepts, 3), or that jump to relevant sections (related concepts, 4). At every step, they can access all research objects indexed with the current keyword (RNA structure' in this example, 2), as well as a list of recommendations from other sources[1]. A dedicated forum can be used to discuss the keyword (its definition, standard terminology, relevance at this location, etc.). A visualisation tool (7, not designed yet) will show how populated terms leading from the current keyword are.

An authenticated user can edit the OWS by adding new keywords (5) or create new horizontal connections to existing inter-disciplinary keywords (6). Newly proposed keywords (8) are subject to an open vote by all authenticated users toward final inclusion in the OWS. Authenticated users can also follow a keyword (9) and get notified of all activity occurring around it (i.e. new keywords connected to it and/or new items tagged with it) in a customizable way (frequency of notification, specific

[1] Many are being developed throughout the Web. There are for instance the self-journals of SJS, the collections of Science Open, knowledge maps, Mendeley groups, etc. OWS aims to also addresses the problem of the fragmentation of recommendation systems; any interesting article will be very hard to miss!

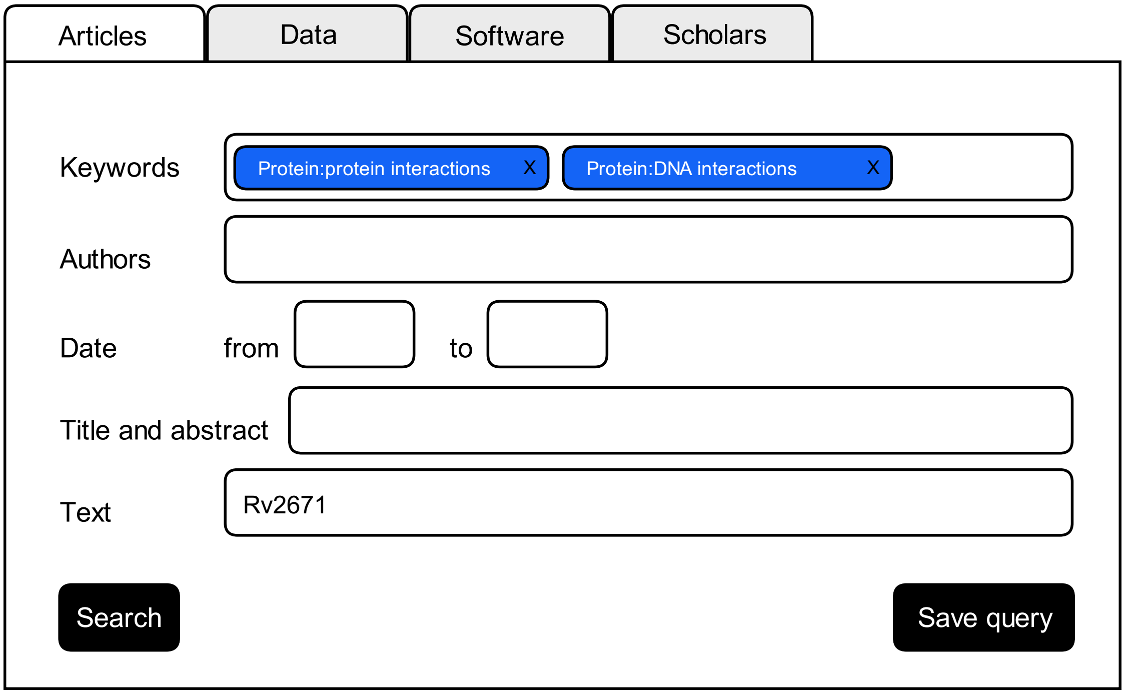
items followed etc).

A responsive search tool with self-completion (10) allows users to quickly center the interface on keywords of interest without navigation.

The OWS will be stored inside a Neo4J database (best suited for graphical relations). The interface will be developed using plain HTML5, CSS3 and JavaScript to provide an ergonomic and responsive experience (which, in particular, is mobile-friendly).

3.2 Advanced queries

Users can also perform complex database queries through advanced search options.



Keywords: Protein:protein interactions X Protein:DNA interactions X

Authors

Date from to

Title and abstract

Text: Rv2671

Multiple keywords can be combined with metadata and text search. The engine relies on Elastic Search and NodeJS technologies so that powerful queries are possible with keywords (e.g. recursively interrogating all the "descendents" of the keyword, or all the keywords connected to it at a certain distance).

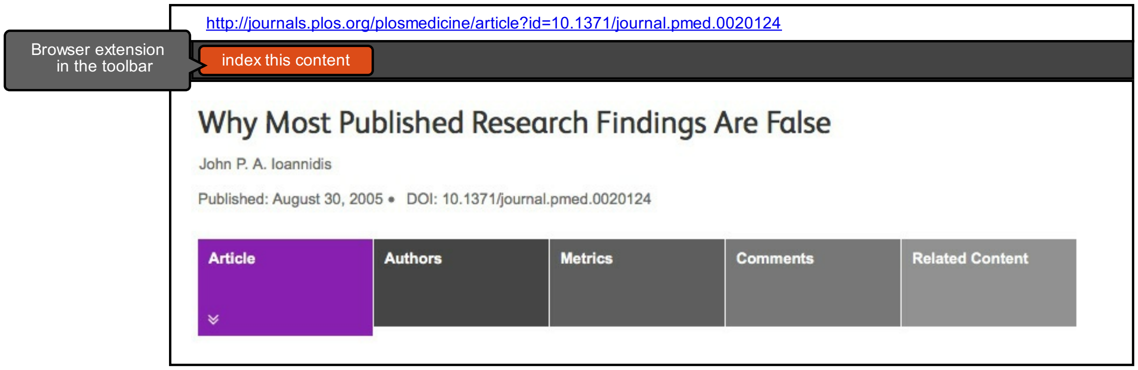
For article searches, results will be sortable according to standard available metrics. The OWS also offers another interesting possibility: by computing distances between the keywords indexing an article, interdisciplinarity can be quantified, and articles sorted accordingly.

Importantly, the development of this user-driven classification system offers an opportunity to experiment with innovative AI algorithms to further expose relations between keywords based on advanced data mining techniques. This possibility will be considered during the design and development phase of the project to facilitate future collaboration with expert AI research groups, as the artificial intelligence workflows that are needed to work from the initial human search constitute a novel research project in itself, beyond the scope of our proposal for the first round of the Open Science Prize.

Finally, the possibility to save queries and follow keywords will enable a periodic notification of new results. Though simple technically, this has the potential to change the way we follow science: All the precise information that users declare an interest in will be delivered automatically. Users only need to pay attention to new keywords and to update their queries as necessary.

3.3 Tagging content

Authenticated users can also tag scientific content with easy-to-use tools in the form of browser extensions which can be either Bookmarklets or plain Javascript browser extensions. They will contain the authentication information and will trigger the indexation interface. This interface consists of a pre-filled form (when possible) with a keyword selection feature. The same process also allows enrichment of articles that have been previously indexed.

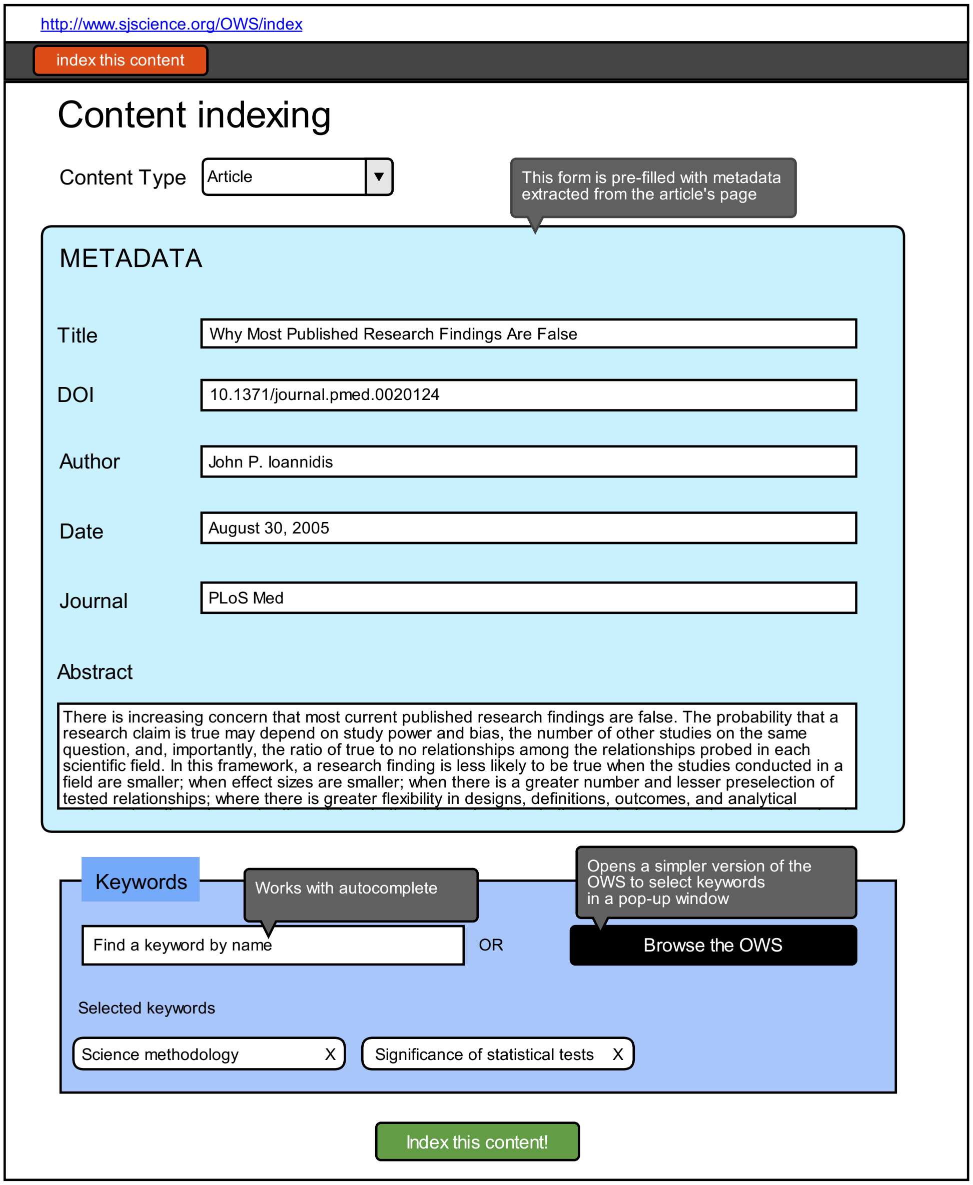

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
Browser extension in the toolbar
index this content
Article
Authors
Metrics
Comments
Related Content

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

When the user clicks the extension button, a new tab opens


http://www.sjscience.org/OWS/index
index this content
This form is pre-filled with metadata extracted from the article's page
Works with autocomplete
Opens a simpler version of the OWS to select keywords in a pop-up window
Browse the OWS
Index this content!

Content indexing

Content Type: Article

METADATA

Title	Why Most Published Research Findings Are False
DOI	10.1371/journal.pmed.0020124
Author	John P. Ioannidis
Date	August 30, 2005
Journal	PLoS Med

Abstract

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical

Keywords

Find a keyword by name OR Browse the OWS

Selected keywords

Science methodology X

Significance of statistical tests X

By tagging their own research (something which will be in their own self-interest to increase its visibility), scholars will be able to effortlessly facilitate the most appropriate classification for the world's scientific output.

Self-moderation mechanism. We have designed a mechanism for the community to moderate inappropriate indexing. This mechanism will not be online initially, and we will only turn to it if problems arise. We will define as correctly tagged anything that is considered to be correctly tagged by more than 2/3 of the community. We can enforce this as follows. Each article/data/software is initially attached to its keywords with a strength of 100%. This strength affects its visibility (i.e. a strength of 0% means that it will no longer appear when the keyword is browsed). Every (authenticated) user can strengthen or weaken this bond but the effect of negative votes is set to have twice the effect of positive votes. Starting from a value of 100%, the strength of the connection between a keyword and the associated digital content stays at 100% if and only if there are twice as many positive votes as negative votes. A potential enhancement of this mechanism would be to also introduce a rewarding mechanism for "good taggers".

3.4 A tool for publishers and repository managers to integrate into their publishing process

When it becomes clear that what we offer is an extremely powerful tool to chart the evolution of science and its developments, publishers will want their content to also be properly indexed. We will offer them JavaScript Widgets that can be integrated into their publishing process, so that new articles/data/software can show up in the OWS as soon as they become available online. We will do the same with database and data repository managers.

3.5 Benefits for open science

The open classification system we propose is a general solution to the problem of knowledge discovery while also directly promotes open science by increasing:

- the awareness, visibility and accessibility of data and software
- the awareness of interconnections beyond pre-determined categories
- community awareness and more importantly, community building.

We strongly believe that Open Science can only really happen in tandem with a cultural shift where scientists start to behave as a community and not as competitors striving to secure a publication slot in a high-impact-factor journal. With our scheme, we create a space to enable scholars to contribute to the collective building of open science in a way that solves many outstanding problems. The OWS will provide an important and positive global precedent for the case of open science.

In particular, the junior scientific community is especially motivated to collaborate openly, to develop their research and contribute to the intellectual discussion within their fields. However, many of the current incentive structures in science create barriers to young scientists being able to contribute and collaborate openly. These barriers most recently came to light as part of an active discussion for young scientists surrounding preprints at the recent #ASAPBio meeting, and this tool fits perfectly with one discussion held at the meeting on allowing developers to create such tools helping users to browse or search for interesting and relevant science. It is through tools such as this that knowledge discovery can promote open science and thus a cultural shift to allow junior scientists to participate more openly in science.

4 Originality of this approach

The problem we are addressing here is well known. So is the idea that collective intelligence should be used to facilitate knowledge discovery. We want to point out the fundamental difference with all other existing approaches. Existing schemes:
(i) generate search results using AI algorithms
(ii) then enrol community suggestions to enrich and/or correct the output of AI-mediated searches.

A general criticism of this approach is that one can never know when, how and by whom, search gaps are filled. While this is not expected to be so detrimental for 'hot topics" where a large and active community provides a watchful corrective eye on the global scale, search quality may be very poor in the vast majority of fields that are less visible but no less important. Crucially, potentially disruptive works are likely to lie in the latter category. The global effort requested from the community might be excessive.

Instead we suggest a scheme whereby:
(i) a community process collectively, intelligently and verifiably defines the categorization of science where all scientific units can be appropriately sorted
(ii) AI can then operate within this space of coherent categories to support more detailed searches along both axes of vertical and horizontal connections.

By allowing collective intelligence to continuously evolve the OWS and then by harnessing the logical search capability of AI, rather than allowing AI to search within an unstructured, post-search, user-corrected space, we are confident of more exhaustive and less noisy results. The workload for the community is also minimal. The main drawback of our approach is that the OWS must be initialized with content before it can generate any useful output. We tackle this issue in the next section.

5 Implementation

In order to generate virtuous community dynamics, motivated individuals must provide the initial driving force. In practice, initializing the OWS means: (i) bringing the classification system to a reasonable level of detail, and (ii) indexing a meaningful fraction of the digital online content (articles, data and software).

5.1 Implementation of the classification system

We will assemble an initial version of the OWS based on an analysis of existing classification systems. To achieve this goal, the resources of choice will include trees that have already been developed by publishers, learned societies or institutions (e.g. PubMed's MeSH, the NCIT, the Dewey-Decimal system and its derivatives, etc.). It is important to point out that these "established" classification schemes have been independently built with a specific purpose in mind. We will therefore develop a methodology to rework and assemble them to best provide an initial set up that the community can then intelligently adjust. Their inadequacies are many: e.g. use of "other" as a category, the merging of separate concepts to save depth, ambiguities (e.g. "homology" as used in biology and mathematics), different thesauri describing the same field with different structures and terminologies, etc. In this regard, it is important to note that we will also utilise output from places where collective intelligence has already been successfully put to use; primarily Wikipedia which has been found to contain more up-to-date connections between keywords, and in particular the "horizontal" linkages between distant fields that are missing in local trees. This will provide our scheme with much additional flavor and categorical power.

In the initial setting of the OWS, our goal is to provide a tree structure whose branches have a vertical depth of at least 4 levels (e.g. Science → Biology → Developmental Biology → all primary keywords of Developmental Biology).

5.2 Indexing content

All online digital content (articles, datasets, software etc) are to be indexed in the OWS. For our scheme to work as proposed, it is important that AI does not perform the interpretation and indexing of content; it has to be achieved in a supervised manner so that as large a volume of quality input as possible can be used to initialize the OWS.

5.2.1 Articles

Our primarily focus will be on journal and repository publications where authors have already provided keywords in their articles. We will map these keywords onto the standardized ones of the OWS and index them accordingly. In addition, we will look for occurrences of the keywords (and their equivalent terminologies) in our OWS and match them with those in the title and abstracts of articles. An occurrence at this level may flag up a potential indexation, however this will be verified with input from the community. Only once we are satisfied that this process is performing as expected, will it be applied on the global scale.

Article items in particular are likely to be under-indexed since authors tend to use very general keywords when tagging their articles in the context of a journal. However, content can be more deeply tagged in a second iteration by users themselves. The addition of less frequent but more specialized indexing keywords is also incentivizing in the context of the OWS since, as a proportion of individual tagged articles, they will have increased visibility. This is a mechanism which will also help the OWS will gain momentum.

All the tools necessary to perform a massive screening of articles are already openly available (e.g. rOpenSci, contentmine.org, etc.). Following application of this important initial step, we plan to make at least 10 million articles accessible from the OWS.

5.2.2 Data & Software

We will engage with specialized database managers and decide collectively the best way to index their data (e.g. globally or by splitting it in relevant parts). We will particularly engage with general data repositories (such as Figshare, Dryad or Zenodo), GitHub, GitLab and Bitbucket, and also offer them a customized turnkey solution to integrate indexation into their submission process.

6 Miscellaneous

License All the code we will write for the OWS will be published with a GPL-3.0 license.

OWS usage data We will also publish regularly the anonymized usage statistics of the OWS (i.e. its volume, how fast it grows, currents associated with individual articles, data sources and software etc, and the frequency of use of each keyword in queries for example). There is potentially a vast treasure trove of data associated with the dynamics of how scientific knowledge evolves.

Development team This project will be developed by Open Scholar CIC in collaboration with Future of Research.

Open Scholar (OS) is a growing, non-profit community of volunteer scholars, librarians, and open science enthusiasts with the mission to to develop ideas and tools that promote open and transparent scientific collaboration for a more fast, efficient and natural organisation, evaluation and dissemination of global knowledge. Since its foundation in 2012, OS has successfully developed two innovative and technically demanding projects: The Open Peer Review Module for open access repositories, and the Self-Journal of Science, where the OWS will be hosted as an independent feature that everybody will be able to use. The organization's working team is comprised of active research scholars with ample theoretical background in alternative scholarly communication models and technical expertise in software development. As a community interest company, OS guarantees that all organisation assets are published under a public license and can never be transferred to a for-profit company or institution.

Future of Research (FoR) is a nonprofit, nonpartisan enterprise whose mission is to help young scientists participate in dialogues that shape the landscape of what tomorrow's research, and researchers, can accomplish. To achieve this, FoR is engaging in activities that include: advocating for future research practices that promote effective and sustainable science; enabling conversations among young scientists on all areas of scientific practice and policy; and addressing issues ranging from proper training to equitable funding and publication models. The aim is to bring benefit to young scientists, to the whole scientific research enterprise, and to the public. FoR activities have illustrated the willingness of young scientists to engage in open science and it is through use and promotion of tools such as OWS and with partners such as OS that FoR aims to reduce the barriers to young scientists participating in an open and collaborative scientific enterprise.