Isidore celebrates its 10th anniversary

Pouyllau Stéphane, Minel Jean-Luc, Capelli Laurent, Bunel Mélanie, Sauret Nicolas, Busonera Pauline, Desseigne Adrien, Baude Olivier, Jouguet Hélène

2021/10/19

Mot-clés : isidore, moteur de recherche, outil de découverte, réseau social académique, réseaux de neurones, latent dirichlet analysis, topic modelling

Keywords: search engine, discovery tool, academic social network, isidore, topic modelling, neural networks, latent dirichlet analysis

Isidore celebrates its 10th anniversary

In 2011, the first version of ISIDORE[1] was published. Its 10th anniversary is an opportunity to recall the history of the project and to present its future outlines, currently being defined within the framework of the "Huma-Num Open Science" program[2]. This year ISIDORE passed the 10 million mark in terms of documents and data indexed, enriched and categorized from more than 9000 databases and data repositories worldwide. More than 2000 researchers already have an account, taking advantage of the many customizable scientific monitoring and discovery features.

So, what is ISIDORE?

Initially, ISIDORE is a search engine for discovering and finding publications, digital data and profiles of researchers in the social sciences and humanities (SSH) from around the world. It allows the full text of several million documents (articles, doctoral theses and dissertations, reports, datasets, web pages, database records, descriptions of archives, etc.) as well as event reports (seminars, conferences, etc.) to be searched. In addition, ISIDORE links these mil-

[1] https://isidore.science

[2] Program funded by the National Open Science Fund of the Ministry of Research, Higher Education and Innovation.

lions of documents together by enriching them with scientific concepts provided by SSH research communities.

It is accessible on the Web on the "isidore.science"[3] web portal.

It also offers scientific social network functionalities. As such, it falls into the category of search engines and assistants and offers many features to organize scientific monitoring.

Founding principles

ISIDORE harvests data and publication repositories in order to retrieve metadata and full text, enrich and index them. The exploitation of document metadata as well as of full text enriches the documents, firstly, by linking them to concepts of scientific controlled vocabularies (thesauri, authority lists, etc.) published either by SSH scientific communities, or by large international research libraries, and secondly, by identifying authors with unique identifiers (ORCID, IDRef, IdHAL, VIAF, etc.). The decision to use the repositories coming from large infrastructures such as national libraries, and to align them with those developed by SSH research teams, puts ISIDORE at the heart of the socio-technical devices of information retrieval. One of the strong points of ISIDORE is the different types of semantic processing performed:

- Annotation or semantic enrichment: the terms present in the document metadata, most often taken from the author keywords, title and abstract, are compared to the repository entries using an algorithm based on a morphological analysis of the terms. If an equivalence is found between a term from the document and an entry in one of the repositories, then the resource will be linked to that repository entry. The repositories are multilingual and aligned with each other. Thus, the semantic annotation is multilingual. Today, ISIDORE enriches in three languages: English, French and Spanish.
- Disciplinary categorization: ISIDORE uses a semantic classifier which, after being trained on a training corpus, categorizes all the documents present in ISIDORE in the 27 SSH MORESS disciplines of the (also used by HAL). The training of the classifier is carried out either by a manual categorization done by the researchers in HAL-SHS when depositing their publications, or by manual categorization of articles entrusted to our colleagues of the INIST-CNRS (in order to balance the training on the three languages of ISIDORE).
- Author detection: ISIDORE detects the authors of documents and enriches the author form (first name and surname) using international (ORCID, VIAF, ISNI) and national (IdHAL, IDRef) author identifiers.

Within this framework, it is quite possible for an SSH research community to

[3] https://isidore.science

propose controlled vocabularies, data, and databases that can enrich the set of documents and data present in ISIDORE. Beyond the web portal isidore.science, well known by research communities, ISIDORE offers several types of access: - For developers, APIs[4] (search engine, controlled vocabularies, suggestions). - For researchers and the world of libraries, documentation and archives, a publication of enriched metadata in the Linked Open Data according to the principles of the Semantic Web (RDF) and through a SPARQL end point[5].

In addition, ISIDORE is used by other portals and sites that exploit certain of its "software components". In 2017, as part of a hackathon with Canadian colleagues from Erudit, the ISIDORE on-demand[6] device was born. This modular device now irrigates other portals such as DARIAH[7] (with for example the experience of the Parallel Semantic Search Engine), STYLO[8] or library discovery tools.

Today, ISIDORE is an ecosystem offering research tools for researchers, doctoral students, librarians and archivists, developers and data-scientists.

A bit of history

The main steps

Designed in 2009, the realization of ISIDORE was led by the team of the Very Large Equipment (*Très grand équipement*) Adonis operated by the CNRS(Maignien 2011) with the assistance of the Atos Consulting company. ISIDORE was developed between October 2009 and December 2010 by the teams of the *Centre pour la communication scientifique directe* (CCSD) with the involvement of Antidot SA, Sword and Mondéca. ISIDORE was launched in "beta" version(Pouyllau 2011) on December 8, 2010, during the Adonis meeting in Valpré (France) and in full version on April 4, 2011[9]. It is currently developed and operated by the Huma-Num infrastructure[10] teams who continuously improve the different parts and technological components.

In 2015, ISIDORE was redesigned for the first time to offer enrichment in 3 languages: French, English and Spanish, using the power of the Semantic Web and the terminological alignment between these 3 languages on thousands of concepts. This has already been done for several decades by national libraries and scientists, in major international repositories such as Rameau (*Bibliothèque*

[4]https//isidore.science/api
[5]https://isidore.science/sqe
[6]https://rd.isidore.science/ondemand
[7]https://www.dariah.eu
[8]STYLO is a Canadian text editor developped by the *Chaire de recherche du Canada sur les écritures numériques* (University of Montréal) and the company *PiNinja* , with the help of Erudit and Huma-Num. See https://stylo.huma-num.fr
[9]https://www.aefinfo.fr/depeche/244085-sciences-humaines-et-sociales-isidore-nouveau-portail-web-30-du-cnrs
[10]https://www.huma-num.fr

nationale de France, France), LCSH (Library of Congress, USA), or BNE (*Biblioteca National de España*, Spain), but also in scientific repositories produced by the international SSH scientific communities. In 2018, a major interface update was carried out in order to internationalize ISIDORE through the implementation of a multilingual interface and to add many features for users. Designed with the direct help of SSH researchers (historians, sociologists, geographers, linguists, etc.) and designers from Atelier Universel company, this new interface, based on the principles of a social network between researchers and a scientific monitoring tool, led to the creation of a new website, isidore.science, which is now part of an international distribution such as, for example, the catalogs of the European Open Science Cloud[11] and Re3Data[12].

Ten years of learning

In 2009, the conception and realization of ISIDORE were based on a few assumptions and convictions of the founders(Poupeau 2016). The first step was to offer a unified access portal to all scientific publications and data produced by the community of researchers in the social sciences and humanities, since in 2009, searching for a publication required access to about ten portals (HAL-SHS, Persée, Cairn, just for France, etc.), each offering a specific search interface. It was then a question of exploiting the possibilities offered by the use of disciplinary thesauri, to offer more powerful search functionalities than those offered at that time by existing search engines (Google Scholar in particular). This is what is called semantic enrichment in the "isidorian" language (see above). This exploitation was concretized by displaying the generic terms proposed by the various thesauri on the Web interface.

There was also a desire to exploit the possibilities offered by machine learning to categorize documents. This learning technique requires a training set composed of several hundred articles for each category in order to build the classifier (mainly using HAL-SHS and collaborations with INIST-CNRS from 2015 for English and Spanish). The final goal was to contribute to the interoperability between the different repositories of scientific publications by clarifying the data model (the ontology) and formalizing it in the languages of the Semantic Web (RDF, RDFS, SKOS, OWL). Over the ten years, these convictions and hypotheses have been confronted with the implacable criticism of users. It became apparent, through discussions with user panels, that the display of enrichments had limitations and could be a source of errors. Semantic enrichment and automatic categorization offer particularly powerful search tools, but their visualization needed to be redesigned, in particular by exploiting multilingualism. This is the first lesson that was implemented in the current version of ISIDORE. It also became clear that the rapid evolution of research topics requires very regular updates of the repositories from the SSH research communities and updates

[11] https://marketplace.eosc-portal.eu/services/isidore
[12] http://doi.org/10.17616/R35V2Q

of the classifier training base. These tasks are extremely time-consuming and could not be performed regularly. A count via a SPARQL query indicates that on July 19, 2021, out of 3,381,711 article-type documents present in ISIDORE, only 2,215,744 are categorized with a confidence of more than 70%, i.e. almost 35% of articles are uncategorized. This second lesson invites us to reflect on the new technologies available to overcome this difficulty. Interoperability, via APIs and the Sparql endpoint, has not led to cross-use with other repositories, particularly on a European scale, whereas other institutions, such as BnF, or projects such as Biblissima, seem to have made better use of these functionalities. The FAIR principles have reinforced the initial choices (the F, A and I) but also questioned the separation between ISIDORE and NAKALA, the latter being in charge of preserving research data (the R). Today, the links between ISIDORE and NAKALA need to be rethought: this is the core of the Huma-Num Open Science program currently underway.

Towards tomorrow...

The multiplication of search and discovery engines since the early 2000s has made the information search landscape (Pouyllau S. 2020) more complex for researchers (Chaudiron 2008). ISIDORE, through its inclusion in 2013 in the Huma-Num ecosystem, benefits however from an integration of digital services (centralized authentication, data repositories created with NAKALA[13] and harvestable, etc.). But while the Web interfaces have flourished, the search functionalities, which rely on the quality of the indexed data, and on the capacity of the socio-technical device to analyze, group, sort and link them, have mostly remained in their initial state. The creation of processing chains in Machine Learning and especially their exploitation require fast and often complex life cycles in terms of choices and scientific validation of enrichment repositories. However, the development of new technologies, such as Deep Learning or Topic Modelling, opens new horizons.

ISIDORE in the Huma-Num eco-system: from continuous improvement to the HNSO program

Since its creation, ISIDORE has evolved regularly, especially in terms of its semantic enrichment workflows and information processing components. Since 2015, with the addition of annotation, categorization and enrichment in English and Spanish, the processing chains have been extensively updated and reworked. While the year 2018 was dedicated to the evolution of the interfaces and the addition of scientific social network functionalities, since 2019, Huma-Num teams

[13]NAKALA, by Huma-Num, is an interoperable and secure service for depositing all types of data (e.g. text files, audio, video, images or other types) in order to share them. See https://documentation.huma-num.fr/humanum-en/

have been working on several improvements that are in progress. We present two of them below, which are currently being finalized.

Managing data in ISIDORE and the arrival of real time in data indexing

Created with a limited number of sources to harvest and index, ISIDORE has grown rapidly in 10 years and the question of real-time updating was quickly raised. The limited resources at the beginning, the difficulties stabilizing the team and other choices such as multilingualism in annotation and categorization have significantly delayed the development of this functionality. For several months, the Huma-Num team, and in particular its ACCES division, which develops the code and components of ISIDORE with Antidot as an industrial partner, have been working on the implementation of a rapid harvesting system: ISIDORE will be updated on a daily basis rather than monthly. The specifications are complex because ISIDORE is not limited to indexing metadata and full text, and the entire processing chain had to be redesigned. In this evolution of ISIDORE, a new database, containing the sources - i.e. the databases to be collected and harvested (in OAI-PMH or with the help of the Sitemap/XML couple and expression in RDFa of metadata) – has been designed by the ACCES division. It will allow data providers to edit their databases, add collection portals, etc.

Knowing the authors better in order to discover their work and establish collaborations

Another ongoing improvement is the detection of document authors. Author identification is a major feature since ISIDORE offers an academic social network layer. Following authors, discovering the journals in which they publish, the works they deposit in open archives, etc. requires the reliability of author detection and of ISIDORE's ability to, for example, disambiguate homonyms and associate the right identifiers (IdRef, Wikidata, BnF, ORCID, VIAF, ISNI, etc.) to the right people. Beyond that, the aim is to offer new functionalities such as suggesting contact between people working on similar subjects but belonging to different countries or schools of thought, or detecting emerging research themes or experts in a very specialized field (Pouyllau, 2020). With the industrial partner Antidot, a new processing workflow has been developed, which allows the analysis of author forms, using machine learning techniques. This is a difficult and meticulous task, which will also require the inclusion of tests with panels of researchers to continuously improve this type of device over the long term.

ISIDORE at the heart of Huma-Num, the work carried out within the framework of the "HNSO" program

More broadly, Huma-Num is committed, with the "Huma-Num Science Ouverte" program financed by the national fund for open science, to developing a vast program of improvement of the two core components: ISIDORE and NAKALA. In this context, a number of bridges have been built between ISIDORE and NAKALA, in particular in the use of repositories and thesauri provided by the international SSH communities. Using ISIDORE repositories to help researchers index their datasets more efficiently in NAKALA, and using the full text of datasets deposited in NAKALA to improve the classification of documents in ISIDORE, are at the heart of the work being carried out at present and until 2024 in the framework of this "Huma-Num Open Science" program. Beyond the interconnections, which are part of the IT aspects of the development, it is also the entire document workflow that must be rethought, in particular at the level of the life cycle (regular updates, etc.) of the scientific repositories and the ontology of ISIDORE types. This work is being undertaken within the framework of this program. With the joint development of HumanID[14], the "hub" for authentication and access to Web services proposed by Huma-Num, a "seamless" integration of services for researchers is being created at Huma-Num and prefigures what will undoubtedly be at the heart of the vast COMMONS project[15], developed in cooperation with the OpenEdition and Métopes infrastructures. Beyond the software components and methods, HNSO is interested in the future of ISIDORE, in the evolution of information retrieval practices of researchers. This is one of the missions of the HN Lab, whose activities, revolving around 4 research axes, feed the HNSO program and thus the development of ISIDORE.

Towards new methods for semantic content enrichment: neural networks or Topic Modelling

In the last 10 years, semantic classification and annotation techniques have made great progress. Two approaches offer a more efficient categorization and less time-consuming processing. First, the recent progress made in the last five years in the development of Deep Learning architectures based on neural networks (CNN, LSTM, Tranformers) and the construction of large coverage language resources (BERT and DocBERT (Adhikari A. 2019), FLAUBERT, CAMEMBERT, etc.) provide categorizations with an F1 score close to 91%. Nevertheless, these approaches still require training corpora that can be complex to produce. Another approach, Topic Modelling or Latent Dirichlet Analysis (LDA), which is based on the construction of a probabilistic model (Blei, 2001) does not require a training corpus. On the other hand, the categories assigned to the articles (the topics) cannot be predefined. The advantage of these methods - for a tool anchored in research communities that work on very different fields and objects - is that they are better suited to the topics produced by the schools of thought. Nowadays, the emergence of these topics, which can

[14]https://humanid.huma-num.fr

[15]See (in French) : https://humanum.hypotheses.org/6466

be seen as "short-lived potential concepts", is accelerated by the very rapid diffusion of articles and books in open access. These topics could be linked to the repositories produced by communities, to the repositories of national libraries already present since 2010 in ISIDORE, or even to scientific definitions coming from cooperative devices such as wikidata, full text article database, etc. For example, the work carried out by the HN Lab on the journals "Intermédialités et Études françaises" in partnership with the DSI Group may contribute to these reflections[16]. As we can see, in terms of semantic annotation chains, several avenues for improvement are open for the next 10 years of ISIDORE.

Chaudiron, M., S. et Ihadjadene. 2008. "Quelles Analyses de L'usage Des Moteurs de Recherche." *Questions de Communication* 14. https://doi.org/10.4000/questionsdecommunication.604.

Maignien, Yannick. 2011. "ISIDORE, de l'interconnexion de données à l'intégration de services." https://archivesic.ccsd.cnrs.fr/sic_00593320.

Poupeau, Gautier. 2016. "Bilan de 15 Ans de Réflexion Sur La Gestion Des Données Numériques." http://www.lespetitescases.net/bilan-reflexion-sur-la-gestion-des-donnees-numeriques.

Pouyllau, Stéphane. 2011. "ISIDORE : une plateforme de recherche de documents et d'information pour les Sciences Humaines et Sociales." https://archivesic.ccsd.cnrs.fr/sic_00605642.

Pouyllau S., Capelli L. et MINEL J-L., Bunel M. 2020. ""We": A Proposal for the Triple Platform," September. Zenodo. https://doi.org/10.5281/zenodo.4059099.

[16]This work is part of the Revue2.0 research program conducted at the University of Montreal. See HN Lab log for more information.