ISTEX: a powerful project for scientific and technical electronic resources archives

To strengthen and facilitate French research: this is the goal the Ministry of Higher Education and Research wants to achieve with its ISTEX project. France, along with the rest of the world, has faced huge price increases in the acquisition of academic electronic resources during the last decade, and its institutions are now finding it difficult to pay to access them. The Ministry has therefore decided to invest in an extensive electronic archives acquisition programme in order to give the academic and research community nationwide access to a very large corpus of data in all disciplines. This project will consist of two phases: acquisition of the archives themselves and the development of a single platform to host the data. This platform will offer seamless unique access to all the resources purchased and will provide a range of efficient services to researchers.

Four partners for a nationwide archives acquisition campaign

ANDRÉ DAZY

Couperin
Coordinator of the department services and forecasting studies
Paris

The main objective of the ISTEX project[1] (initiative for excellency in scientific and technical information) is to offer to the entire higher education and research community remote online access to the retrospective collections of scientific literature in all disciplines. This national archives acquisition policy will comprise the purchase of mainly journals archives, databases and corpus of texts.

This project is part of the 'investment for the future' programme initiated by the Ministry of Higher Education and Research, whose ambition is to strengthen the position of French research and higher education on the global scene. Four partners lead the initiative: the National Centre for Scientific Research (CNRS)[2], the National Bibliographic Agency (ABES)[3], the consortium Couperin[4] and Lorraine University[5] (on behalf of the University President Conference). Signed in April 2012 by the State, the National Agency for Research (ANR)[6] and the CNRS, the convention has a €60 million budget for three years: €55 million for the acquisitions themselves, of which €32 million has already been spent, and €5 million to build the platform that will host the data.

Two-level governance

The four partners together form the executive committee, assisted by a representative of the Ministry, and they are in charge of overall coherence and of the follow-up of the project. This committee is accountable to the steering committee composed of representatives of the universities, the research institutions, the Grandes Écoles and the Ministry, all in charge of policy responsibility in the IST field, and responsible for the political guidance of the project. The steering committee validates the decisions taken at the executive committee level, lists the final resources that will be negotiated, ensures that a balance is reached between the various research fields and that all the higher education and research communities' interests are taken into account. When no majority can be obtained, it also has the casting vote in deciding certain issues.

A big survey to fit the project to researchers' needs

A nationwide survey[7] was launched in 2012, prior to the acquisition stage, to better understand researchers' needs. Some 7,167 researchers (representing 7.5 % of French researchers) responded, and proposed 1,648 different publishers and 5,624 resources. At the same time, an invitation to tender was announced, resulting in a total of 236 offers. After cross-reference with the researcher survey results, 25 resources were selected.[8] Negotiations then began, starting with a target price calculated by Couperin and the ABES and with the collaboration of Jisc for a few resources. The first contract was signed in December 2013. A nationwide consultation was undertaken in 2013 regarding the resources that had been tendered, some of which were of specialized interest or previously unknown in France. This enabled the French research community to evaluate the 236 offers and assess their scientific interest and utility. The consultation also raised researchers' awareness of the cost of these resources, of which they would otherwise have been unaware. In the meantime, the platform has been in development by the Institute for scientific and technical information (INIST)-CNRS[9].

> "… an exceptional corpus of several million multilingual and multidisciplinary documents via a single platform."

A powerful platform providing helpful services to the research community and librarians

The second phase of the project consists of creating the platform which will host the resources that have been procured. As mentioned, INIST-CNRS is developing the platform and this will be available in 2015. Until then, access to the resources is via the publishers' platforms. ABES manages the national licences to the resources and enables access via a dedicated website[10], which also provides academic libraries with practical information about contracts (licence terms) and metadata. An application hosted on this website[11] allows institutions to create an account and state their IP addresses, which are then forwarded to the publishers.

An open window on a unique and exceptional corpus

ISTEX will be a unique tool, the first one to provide an exceptional corpus of several million multilingual and multidisciplinary documents via a single platform.

A systematic access to the full text

The platform will not be a repository of metadata linking to documents hosted by publishers, but a database gathering all the full-text content from a diverse typology: journal archives, archives or heritage resources and databases, etc. This will enable different but complementary use, independent of external publishers' access authorization, without time limitation and with the possibility of text and data mining (TDM), either the whole database or facets of it according to disciplinary needs or by search criteria such as date range or document type.

A powerful search engine adapted to researchers' needs

A powerful search engine adapted to researchers' needs and with easy-to-use search and upload facilities, data treatment, data extraction and TDM, will give life to the platform. The first step is to check that the metadata supplied by publishers are accurate and that their quality is TDM compliant. The search engine has to be adapted to comply with the high standard required to meet the needs of scientific research. This includes an automated language treatment tool for multilingual content and a lemmatization module. ElasticSearch[12], an open source search engine, has been chosen. This will enable the platform to benefit from developments made by the user community. This means that this huge repository will be suited to applied research such as history of sciences, and will be useful for documentary synthesis, terms extraction, literature review, semantic ontology and metrics. Furthermore, it will be totally integrated in the national documentation

landscape and will allow exchanges with other projects in the same area such as resources management, for example.

Specific services under development

The basic service will enable search across articles and collections and full-text indexing. Other services under development will provide for deeper search of full text. Two teams from LINA[13] and INIST-CNRS are currently working on the detection of terms and their variant spellings in the full text and on a scientific terminology repertory for ISTEX data exploitation.

Named entities extraction A research team from the computer laboratory of Tours and INIST-CNRS is in charge of developing a program to detect, standardize and tag dates, names, town, region, country, family, research team, research project, laboratory, institution, resource internet addresses and the names of the stars, molecules, mathematic formulas, plants, etc.

Access to the main fields of the bibliographic references INIST-CNRS is undertaking automated tagging. This work will allow researchers to build scientific maps and to answer queries such as: Who is working with whom? Which are the existing network citations? Where do researchers publish? Will their research evolve over time?

Three advanced services will also be available:

- *CILLEX project*: led by the CLLE Toulouse[14]. Work is under way on a search engine with automated classification response. The project aims to develop metrological tools to make it possible to identify the relevant information. The results of a search in ISTEX will undoubtedly result in a large number of references, which will need classification to be usable.
- *ISTEX-R project*: the LORIA[15]. The ATILF[16] and INIST-CNRS are working on this project, the objective of which is to analyse the content of ISTEX and, by means of diachronical maps, to measure the evolution of the research and the stock of knowledge through time.
- *LorExplor project*: aims to construct an open source library of XML components to enable the exploration of the ISTEX corpus. This will facilitate the work of librarians in building (in a few days) intermediary regional or thematic or institutional platforms (from 100,000 to one million documents) or in answering specific queries.

A platform integrated in the local tools

The platform will allow easy connection to existing portals and discovery tools or link resolvers, including commercial ones, via the application programming interface (API) or widgets, or OAI-PMH harvesting, and will be able to easily plug in the content management system (CMS) used by libraries to create seamless access to both archives and current subscriptions.

Remote access for all

Remote access will be available for all the members of the institutions of the Higher Education and Research Ministry and at some public libraries as well.[17] Access will first be enabled by IP addresses – 254 institutions already have access to the first ISTEX negotiated resources – and then by authentication. A demonstrator portal with a browsing interface will be developed as a solution for those institutions that do not have their own CMS.

"A long-term programme will secure the data for decades."

Eventually, this platform will also be connected to HAL[18], the French open access (OA) repository, which will allow access to OA publications, giving them greater visibility. Connections to some other European repositories are also being considered.

A long-term programme will secure the data for decades. The National Computing Centre for Higher Education (CINES)[19] is in charge of this preservation.

Many advantages for all

This centralized acquisition policy presents many advantages for all the stakeholders. ISTEX allows for an equality of access across all districts and institutions in France. All the users will have access to the resources regardless the institution they belong to, large or small. The platform will cover all the scientific fields, users will have at their disposal a multidisciplinary repository, which will permit collaborative research across institutions. The content provided by the platform will complement the current journal content to which institutions subscribe. The enhanced bargaining power enabled through acquisition of content by a national licence will save public money, and national negotiation has ensured that TDM is a non-negotiable requirement and must be permitted if publishers wish to be successful in their tender submission.

> "… a fruitful synergy to the benefit of the community …"

The project is still at an early stage so there has not yet been any user feedback, but there have already been beneficial outcomes for the IST community. For example, the negotiations have raised issues about TDM and IPR in derived data, with the view emerging that TDM should be permitted by publishers as a matter of course and not as an optional feature.

By working closely together on this project, the four main partners have created a fruitful synergy to the benefit of the community they serve.

References

1. ISTEX: http://www.istex.fr/?Presentation (accessed 18 August 2014).
2. CNRS: http://www.cnrs.fr/index.php (accessed 16 August 2014).
3. ABES: http://en.abes.fr/ (accessed 17 August 2014).
4. Couperin: www.couperin.org (accessed 17 August 2014).
5. Université de Lorraine: http://www.univ-lorraine.fr/ (accessed 17 August 2014).
6. ANR: http://www.agence-nationale-recherche.fr/en/project-based-funding-to-advance-french-research/ (accessed 17 August 2014).
7. National survey: http://enquete.inist.fr/launay/istex/istex.hyp (accessed 16 August 2014).
8. Selected resources: http://www.istex.fr/ressources-selectionnees-en-2013/ (accessed 16 July 2014).
9. INIST-CNRS (NB: CNRS is the official name; the INIST is a part of the CNRS): http://www.inist.fr/?lang=en (accessed 16 July 2014).
10. National licences: www.licencesnationales.fr (accessed 17 August 2014).
11. National licences access: http://acces.licencesnationales.fr (accessed 18 August 2014).
12. Elasticsearch: http://www.elasticsearch.org/ (accessed 10 August 2014).
13. LINA: https://www.lina.univ-nantes.fr/?lang=en (accessed 18 August 2014).
14. CLLE Toulouse: http://clle.univ-tlse2.fr/ (accessed 16 August 2014).
15. LORIA: http://www.loria.fr/loria-news?set_language=en (accessed 16 August 2014).
16. ATILF: http://www.atilf.fr/ (accessed 17 August 2014).
17. Licences nationales: http://www.licencesnationales.fr/bibliotheques-publiques/ (accessed 16 August 2014).
18. Hyper Articles en Ligne: http://hal.archives-ouvertes.fr/index.php?langue=en&halsid=flvhirmq0pkjs9r4evmojrdn74 (accessed 16 August 2014).
19. CINES: https://www.cines.fr/en/ (accessed 10 August 2014).

André DAZY
Couperin
Coordinator of the department services and forecasting studies, Collège Sainte Barbe, 4 rue Valette, 75005 PARIS
Tel: +33(0)156817680 | E-mail:andre.dazy@couperin.org

ORCID iD: http://orcid.org/0000-0002-1353-3493

To cite this article:

Dazy, A, ISTEX: a powerful project for scientific and technical electronic resources archives, *Insights*, 2014, 27(3), 269–273; DOI: http://dx.doi.org/10.1629/2048-7754.157