D7.1 Annex 2

State-of-the-art open access research data for the Humanities in FRANCE

1. The French Open Access Research Data Ecosystem

1.1 Introduction

The CNRS in connection with the University of Aix-Marseille has implemented an ecosystem which aims to cover the entire life cycle of the production of scientific data and publications. This ecosystem is based on the following organizations:

- Open Editions offers comprehensive services in journal publications, books, scientific blogs and scientific events in open access. (Https://www.openedition.org/).
- CCSD (Centre pour la Communication Scientifique Directe) offers a set of services for the management of open archives (HAL SHS) (https://halshs.archives-ouvertes.fr/)
- The Very Large Facility TGIR Huma-Num offers a range of services dedicated to the production and reuse of data

In this report, we focus on the tools offered by Huma-Num that fall within the scope of WP7.1 and WP7.2.

The workflow implemented by Huma-Num has been built on interoperability. The aim is to foster the exchange and dissemination of metadata but also of the data itself via standardized tools and lasting, open formats. These tools, developed by Huma-Num, rely on Semantic Web technologies, mainly for their auto-descriptive features and the opportunities for enrichment that they provide.
The first objective is to promote the sharing of data so that other researchers and communities can reuse the data in an interdisciplinary perspective, and if need be with other methods. For example, a map is a scientific object that can be analyzed and described from the point of view of a geographer or that of a historian. More generally, these services which rely on the principles and methods of the Semantic Web (RDF, SPARQL, SKOS, OWL) make it possible to document or re-document data for various uses without locking them in inaccessible silos. Other interoperability technologies complete them, such as the OAI-PMH. Another important point is to make the storage of data independent of the device that displays and reports the data.
The second objective is to prevent the loss of data by preparing their preservation over the long term. The documentation associated with appropriate formats, which are the basis of data interoperability, greatly facilitates the archiving process.

All the services (see Figure 1) developed by the Huma-Num accompany the life cycle of research data and are designed to meet the needs of scholars in Humanities and Social Sciences:

- SHARE and DISPLAY. The Nakala tool offers services to store, display and share documented and standardized metadata and data based on interoperable technologies;
- DISSEMINATE. The NAKALONA tool in connection with the CMS Omeka provides means to editorialize data stored in Nakala and offers the features of this CMS such as its search engine;
- TAG and PUSH. The ISIDORE tool enriches the data stored relying on several disciplinary thesauri and provides the functionality of a faceted search engine in order to ensure better visibility.

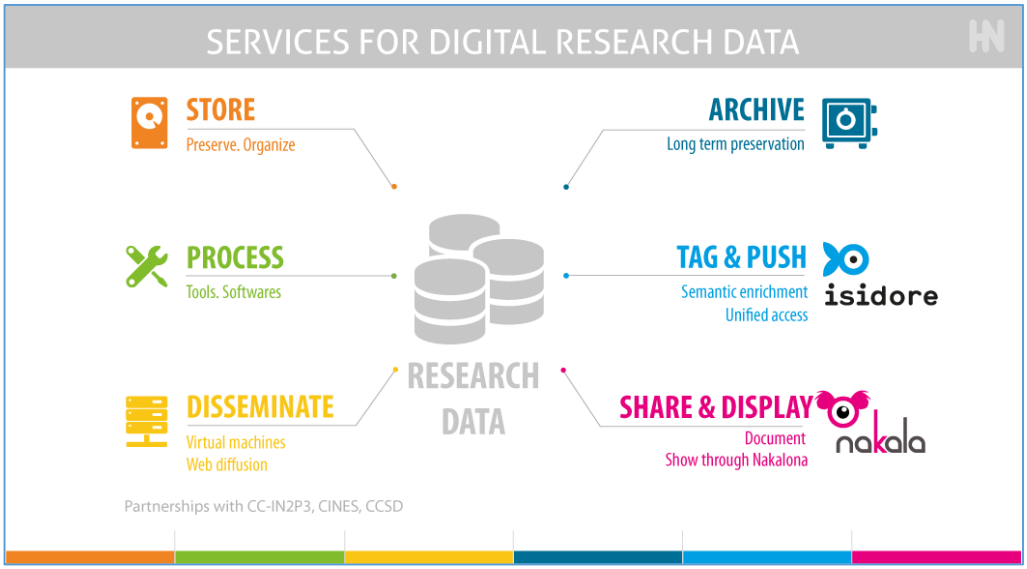


Figure 1 : French services for digital research data

2. NAKALA: SHARE and DISPLAY

Noting that many teams and research projects do not have the necessary digital infrastructure that will provide a persistent and interoperable access to their digital data, Huma-Num has implemented a service called NAKALA exposure. NAKALA offers two types of services: one to give access to the data and another one to expose metadata. By relieving scholars of technical management, it enables them to concentrate on the scientific value of their data. Data hosted by Nakala may be:

- Editorialized with the NAKALONA pack (combining Omeka and Nakala) developed and managed by Huma-Num;
- Shared via a SparqlEndPoint or via the OAI-PMH protocol;
- Searched via the multilingual and multifaceted access platform ISIDORE (see below).

Figure 2 shows these different possibilities.

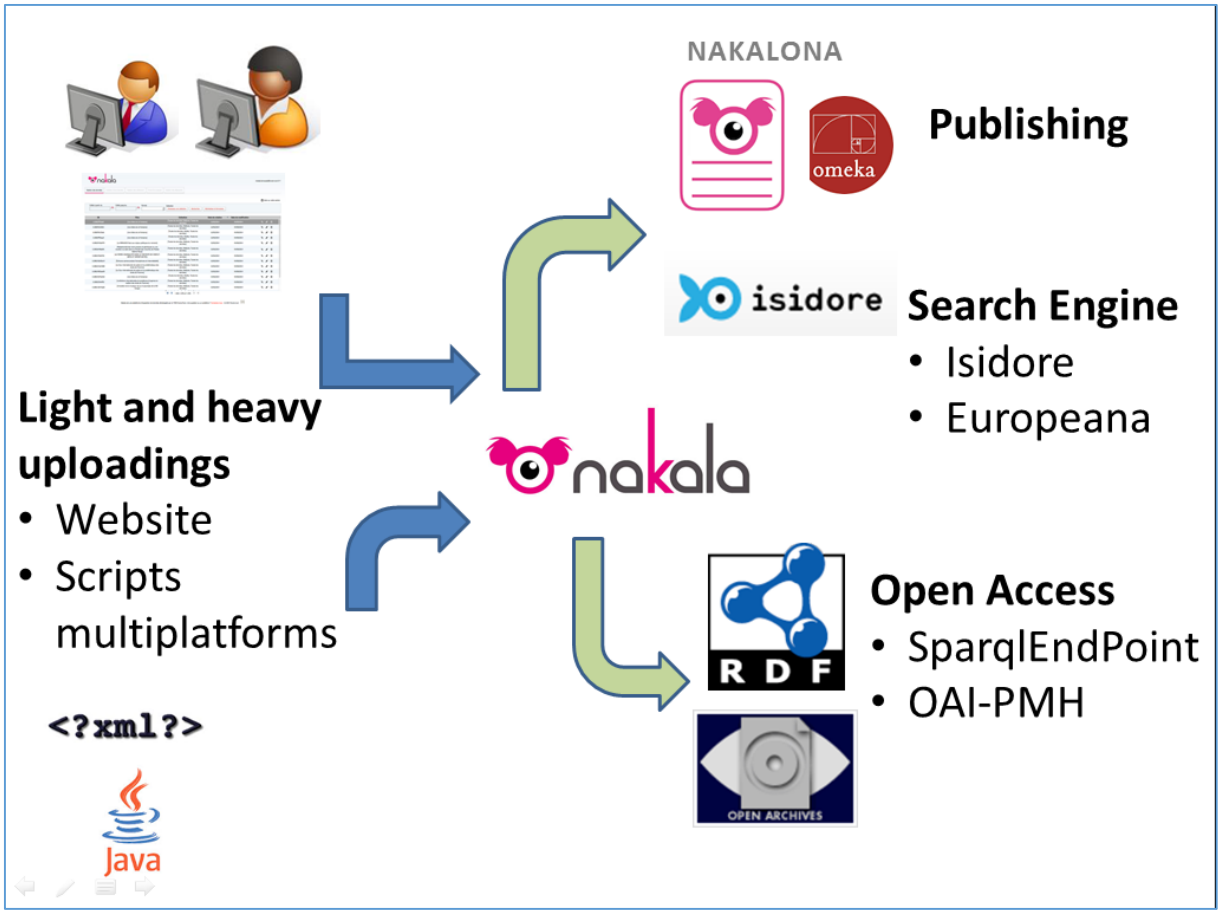


Figure 2 : Principles of NAKALA

2.1. NAKALA Functionalities

- *Persistent identifiers PID*

A unique identifier is associated with each record; data can therefore be quoted regardless of their actual location. The proposed technology is based on assigning handle type identifiers that have a generic addressing mechanism, unrelated with the institution that carries the service. This gives long-term preservation access to data even in the event of change or evolution of the institution that carries the service. The technology used is the Handle system.

- *Permanent access*

Permanent access to data is ensured by the use of the infrastructure of the Huma-Num grid.

- *Data security*

Data are stored on the infrastructure managed by the Huma-Num and are backed up regularly. Huma-Num owns its secure servers within the data center of the IN2P3-(National Institute of Nuclear Physics and Particle Physics of the CNRS), a partner since 2003, and takes advantage of the IN2P3 backup infrastructure.

- *SparqlEndPoint*

NAKALA provides a SparqlEndPoint which enables the sharing of data using the principles and methods of W3C. Description metadata in RDF format are based on vocabularies considered as standards in the community (Dublin Core, Dublin Core extended, FOAF, ORE, SKOS, RDFS and VCARD2006; see Figure 3 and Table 1).

- *OAI-PMH Protocol*

NAKALA provides each data producer with their own warehouse compatible with the OAI-PMH protocol. Since 1999, OAI-PMH is the metadata interoperability standard used in the field of research and higher education. The data can be "harvested" by specialized services, such as ISIDORE, EUROPEANA, OPENAIRE and so on, which increases data visibility.

2.2. NAKALA Access

The services offered by NAKALA are free and in open access. The authentication service used is the French Identity Federation (RENATER), which allows scholars to connect with their login issued by their institutional affiliation (CNRS, University, etc.).

- *Management interface*

The management interface gives an overview of the space used and provides means to manage data and collections. This interface can also be used to upload limited datasets (< 100 items).

- *Management of large files*

To upload large datasets, NAKALA provides a batch processing system. This tool developed in Java can run on almost any platform.

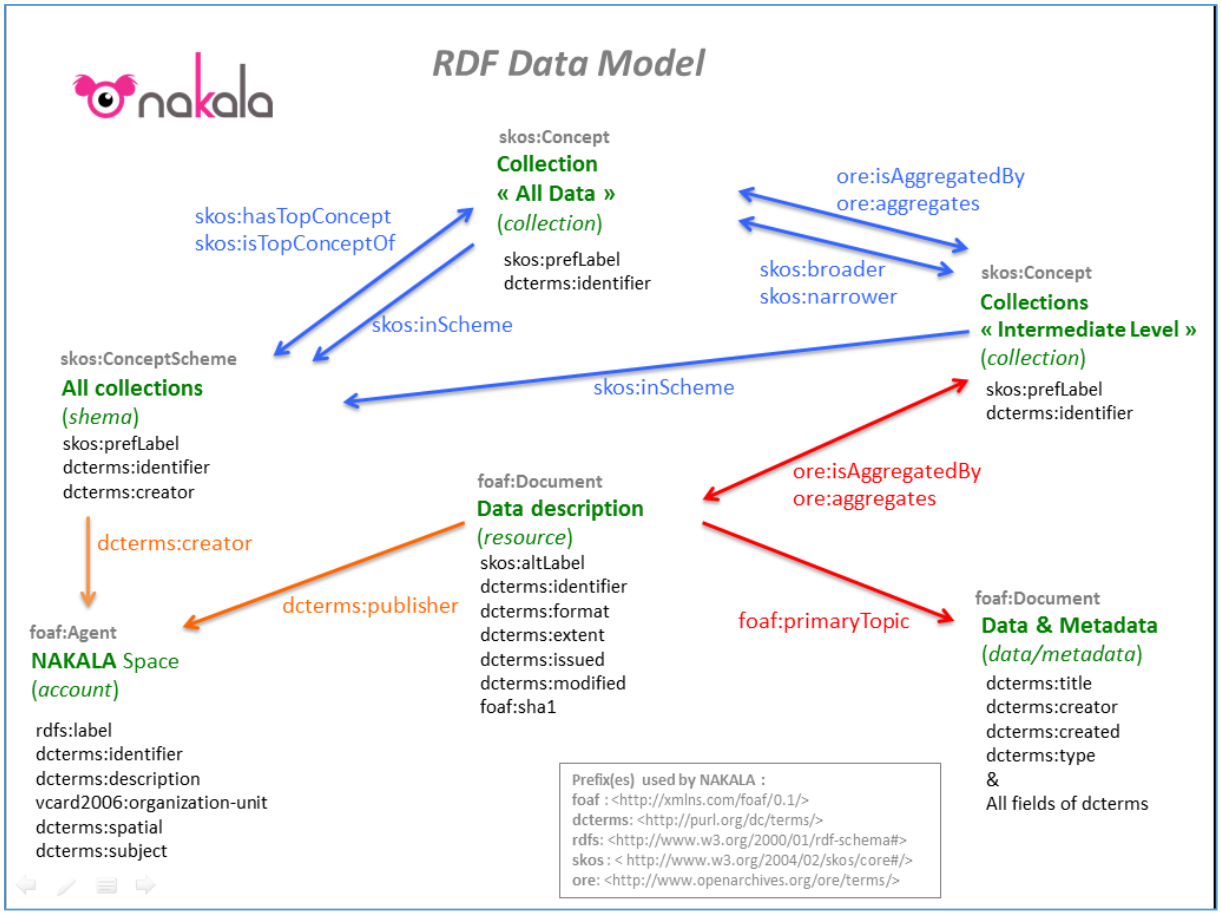


Figure 3 : NAKALA workflow

2.3. Volume

NAKALA was launched in Sept. 2014. So far, 45 data collections are stored in NAKALA, which means 50 000 files for 300 Gb. Since October 2015, deposits have grown by 200 %. An example of a portal built over NAKALA presenting 12 000 pictures of Walls from the Egyptian site of Karnak can be viewed at: http://www.cfeetk.cnrs.fr/archives/

2.4. NAKALA in a Nutshell

The key points of NAKALA are the following:

- W3C standard languages used;
- Secure servers located in Europe;
- Sustainability. NAKAL is managed by the Very Large Facility Huma-Num which is a unit of the CNRS, guaranteeing continuity of service;
- PID management based on Handle;
- Open source development based on reliable and proven components (Handle server, PROAI OAI PMH server and Virtuoso Triple Store server)

2.5. NAKALA Vocabulary

Table 1: vocabularies used in NAKALA

VOCABULARIES	PROPERTIES	COMMENTS
Dublin Core (dcterms)	title	mandatory
	contributor	optional
	coverage	optional
	creator	mandatory
	created	mandatory
	date	optional
	description	optional
	format	optional
	identifier	optional
	language	optional
	publisher	optional
	type	mandatory
	rights	optional
	source	optional
	subject	optional
DC qualified(dcmi)	abstract	optional
	accessRights	optional
	accrualMethod	optional
	accrualPeriodicity	optional
	accrualPolicy	optional
	alternative	optional
	audience	optional
	available	optional
	bibliographicCitation	optional
	conformsTo	optional
	dateAccepted	optional
	dateCopyrighted	optional
	dateSubmitted	optional
	educationLevel	optional
	extent	optional
	hasFormat	optional
	hasPart	optional
	hasVersion	optional
	instructionalMethod	optional
	isFormatOf	optional
	isPartOf	optional
	isReferencedBy	optional
	isReplacedBy	optional
	isRequiredBy	optional
	issued	optional
	isVersionOf	optional

	license	optional
	mediator	optional
	medium	optional
	modified	optional
	provenance	optional
	references	optional
	relation	optional
	replaces	optional
	requires	optional
	rightsHolder	optional
	spatial	optional
	tableOfContents	optional
	temporal	optional
	valid	optional
FOAF	Agent	
	primaryTopic	
	sha1	
ORE	aggregates	
	isAggregatedBy	
SKOS	altLabel	
	broader	
	hasTopConcept	
	inScheme	
	isTopConceptOf	
	narrower	
	prefLabel	
RDFS	label	
VCARD2006	organization-unit	

3. ISIDORE: HARVEST and SEARCH

ISIDORE is a platform allowing access to digital data in the Humanities and Social Sciences. Its architecture relies on the languages of the semantic web (RDF/RDFS/OWL) and provides open access to data. ISIDORE is a managed by the Very Large Facility TGIR-Huma-Num (CNRS, Aix-Marseille University, Campus Condorcet) and implemented by the Center for Direct Scientific Communication (CCSD / CNRS).

The key points of the ISIDORE platform are the following:

- **Targeted harvesting** of metadata and scientific data structured according to international standards available in open access;

- **Indexing of unstructured data** (full text of a scientific article, for example) and of structured data (documentary metadata, for example);
- **Standardization** of metadata and enrichment of data relying on vocabularies recognized in the community (DC, DCterms, FOAF, ORE, RDFS, SKOS);
- **Multilingual (English, French, Spanish) search GUI** exploiting the richness of structured data and vocabularies to make the user an actor of his search;
- **SparqlEndPoint** on sources and indexed data;
- **Smartphone** applications;
- **Supplying** metadata enriched by several multilingual thesauri;
- **Possible integration of the search engine Isidore** in another environment by providing API and widgets.

The following figure shows the architecture of ISIDORE

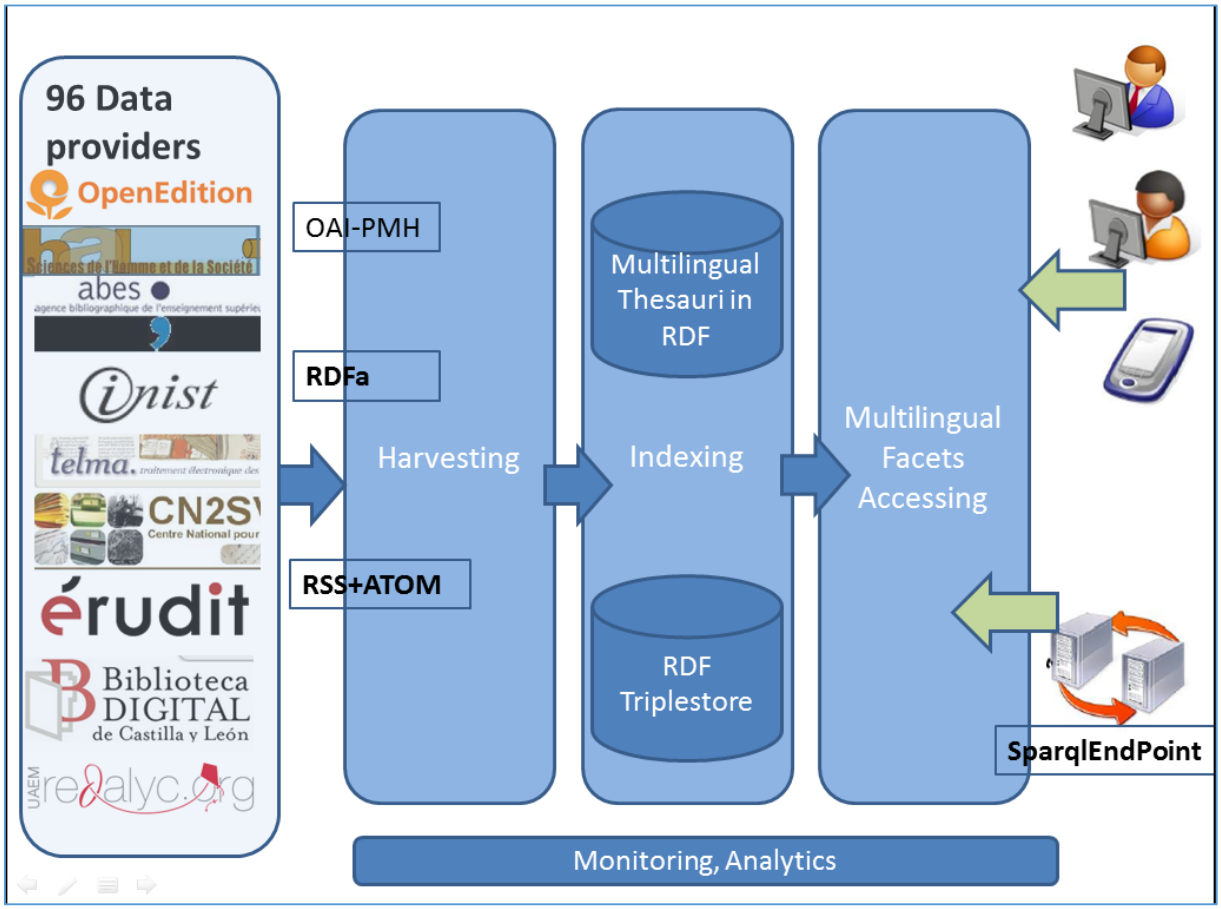


Figure 4 : General architecture of ISIDORE

Here are some details about the key points.

3.1. Targeted Harvesting

So far 3559 sources of data and metadata have been harvested. To be harvested, data providers submit a proposal briefly describing their data and agree to the following conditions:

- Providing OAI-PMH, RDFa or RSS access to their repository;
- Respecting standard descriptions using the following vocabularies (Table 2)

Table 2: Vocabularies used to harvest data providers' warehouses

VOCABULARIES	PROPERTIES	COMMENTS
Dublin Core (dcterms)	*	All properties
DC qualified(dcmi)	*	All properties
RSS	*	All properties
ATOM	*	All properties

The harvesting can be incremental or full. Presently, ISIDORE carries out a full harvesting once a month and several incremental harvests a week. It must be pointed out that this process provides several analyses of the data harvested, especially concerning standardisation errors (date, language, names, etc.). These analyses are provided to the data provider, which contributes to the enhancement of their data quality.

3.2. Indexing and enrichment of harvested data

First, harvested data are structured by using standard languages such as Dublin Core (cf. Table 2) and indexed. Then, ISIDORE uses the following vocabularies and thesauri to enrich the harvested data:

- the HAL vocabulary (Open Archives in Humanities and Social Sciences); the scientific disciplines available in HAL for SHS are from a European vocabulary, the European University Association (EUA) built under the project MORESS - Mapping of Research in European Social Sciences and Humanities. This is a simple and readable nomenclature for researchers, to improve access to information on research in the social sciences and humanities.
- the index of the thematic categories of Open Edition. This index is composed of thematic categories covering all the arts and the human and social sciences. It is developed in the context of electronic publishing platforms and scientific communication OpenEdition's portal. Designed to represent the richness of topics and objects of human and social science research in a single hierarchical index, it is divided into four main categories: societies, mind and language, periods, spaces.
- The RAMEAU vocabulary (Unified Directory of subject, encyclopaedic and alphabetical authorities: Bibliothèque Nationale de France) is a language indexing material that covers all areas of knowledge and applies to all types of papers on all types of media. This documentary language is used, in France, by the Bibliothèque Nationale de France, university libraries, and numerous public or research libraries.

The core of the RAMEAU authority record is trained by common nouns (approximately 100 000) and geographical names (approximately 50 000). It is a controlled hierarchical vocabulary that associates themes by semantic relations (generic terms / specific / associated). The concepts of RAMEAU were converted into RDF SKOS language, as part of the European project TELplus. Each concept, designated by a permanent URI, is provided with labels (preferred or alternative), diverse notes, but also semantic links to other RAMEAU concepts (generic concepts, associated concepts) and towards external repositories (LCSH, DNB). This vocabulary is maintained up to date on the site data.bnf.fr.

- the thesaurus PACTOLS. PACTOLS (acronym of Peoples, Anthroponyms, Chronology, Place-names, Works, Places and Topics) is a thesaurus specialized in archaeology and in sciences of the Antiquity. Archaeology covers the period from Prehistory to World War II, and includes all the sciences necessary to the study and conservation of its objects: human paleontology, natural sciences, physics and chemistry, etc. Another area concerned by PACTOLS is the sciences of Antiquity: from the invention of writing to the year one thousand, in all its aspects. PACTOLS is a poly-hierarchical thesaurus consisting of six micro-thesauri, multilingual (French base translated into Arabic, Dutch, English, German, Spanish and Italian), scalable and autonomous.
- GEMET, the GEneral Multilingual Environmental Thesaurus, was developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA). GEMET was designed as a "general" thesaurus that aimed to define a common general language and a core of general terminology for the environment.
- GeoEthno is a geographical thesaurus designed for geographic indexing of documents in the field of ethnology. Currently under development in the library Eric de Dampierre of the Laboratory of Ethnology and Comparative Sociology, it is used for indexing and querying the library's database and more broadly the ethnology network database. Its coverage is irregular and not exhaustive. It contains about 15,000 terms. It is built around a list of names of countries and territories (ISO 3166-1: 1997)
- The GeoNames geographical database contains over 10 million geographical names and over 8 million unique characteristics, 2.8 million populated places and 5.5 million other names. GeoNames integrates geographic data such as place names in different languages. All lat / long coordinates are in WGS84 (World Geodetic System 1984).
- The Library of Congress Subject Headings (LCSH). By virtue of cooperative cataloguing other libraries around the United States also use LCSH to provide subject access to their collections. In addition LCSH is used internationally, often in translation. LCSH in this service includes all Library of Congress Subject Headings, free-floating subdivisions (topical and form), Genre/Form headings, Children's (AC) headings, and validation strings for which authority records have been created. The content includes a few name headings and geographic headings that are added to LCSH as they are needed to establish subdivisions.
- The Biblioteca Nacional de Espana (BNE) catalogue. The data are derived from bibliographic catalogues and authorities of the National Library of Spain. The data corresponding to certain elements (authors, subjects, works) are enriched with

links to their equivalents in other data sources. For the authors, links are provided, if available, with the Library of Congress, the National Library of Germany, the National Library of France, Sudoc, the National Library of Sweden, VIAF and ISNI. The subject, geographical and genre/form authorities are aligned with their equivalents in the Library of Congress. Isidore uses only National Library of Spain subject authorities available in SKOS.

It should be pointed out that the process of enrichment of a document is carried out according to the document language, in other words French documents are enriched using French (or multilingual) thesauri, English documents using English thesauri and so on.

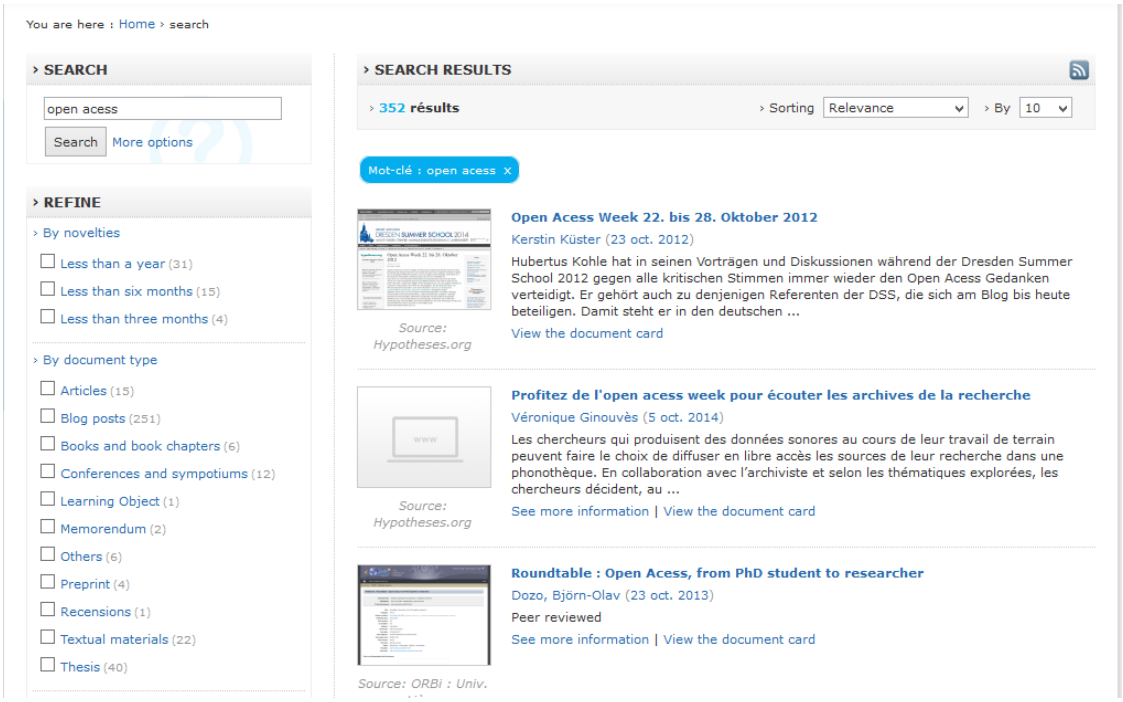
The re-exposure of the enriched metadata follows, in turn, the principles of Web of data by supplying open access to public data. As such, the ISIDORE platform contributes significantly to this movement for research in social sciences in France.

3.3. PID

If the harvested resource has a persistent identifier assigned by an authority (DOI, Handle ARK, etc.), then this identifier is recovered and used by ISIDORE. If the resource does not have a persistent identifier, the system creates an identifier using the Handle system.

3.4. Multilingual (English, French, Spanish) Graphical User Interface

Since 2015, ISIDORE offers a multilingual graphical user interface (GUI) in English, French and Spanish. Figure 5 shows the GUI in three languages for the query “open access”.



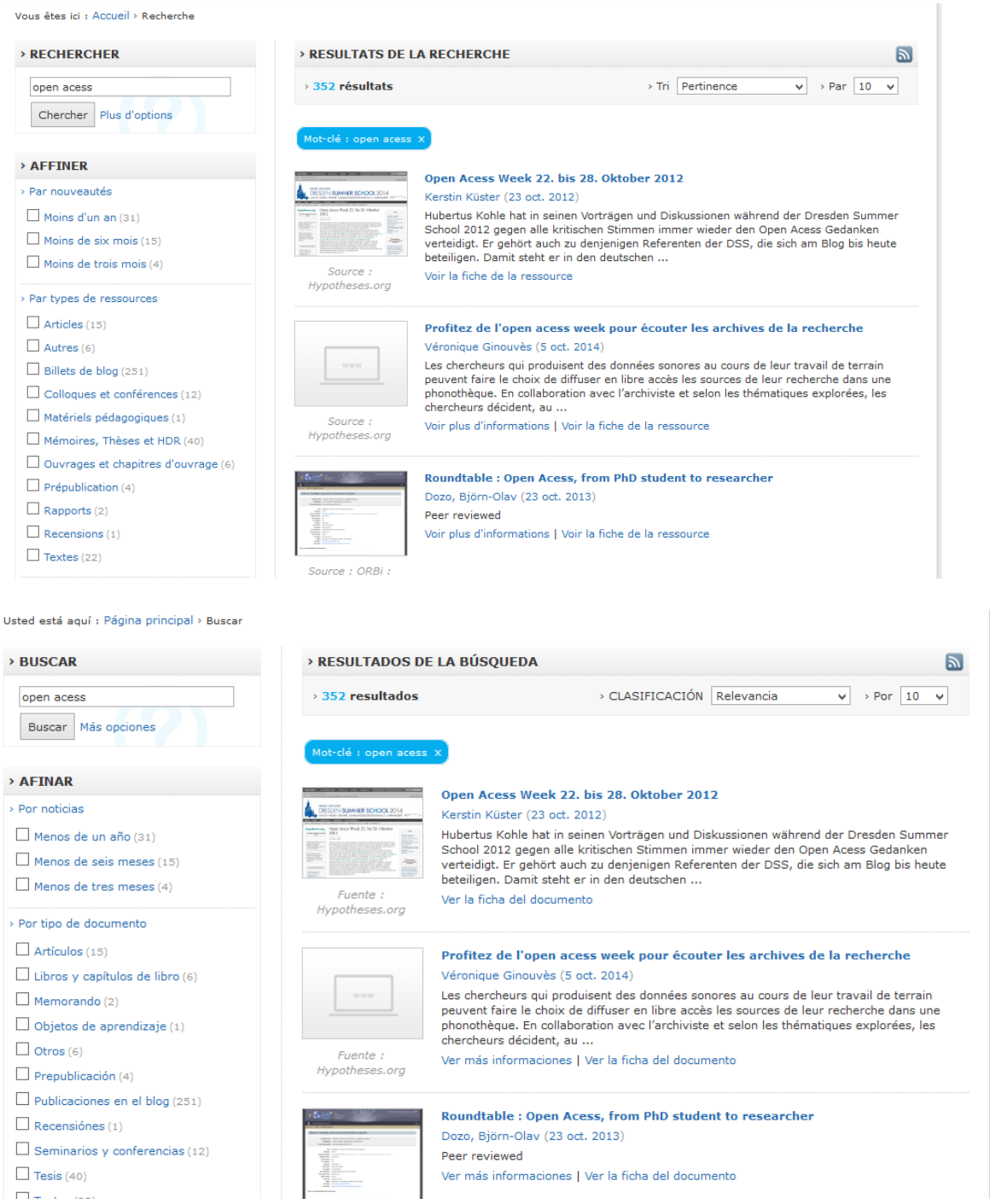


Figure 5 : Multilingual and multifaceted access

Facets offer means to refine a query by using criteria such as freshness of data, document type, categories, historic periods, disciplines, languages, nature of sources and data providers. After selecting a document, it is possible to access all the documents by the same author, to bounce to search for documents on a similar theme or in a similar discipline. Furthermore, ISIDORE suggests other reading based on the database of past queries (readers who read this paper also read these papers).

3.5. SparqlEndPoint and Interoperability

A SparqlEndpoint based on the Virtuoso software offers the possibility to express queries using SPARQL languages. Result formats include HTML, XML, JSON, CSV, RDF/XML. In 2013, the TGIR Huma-Num and DANS developed a prototype (Proof of Concept) in order to show the connection between two repositories NARCIS (DANS) and ISIDORE. The connections relied on the two SparqlEndpoints and the alignment of the disciplinary vocabularies used by these systems (cf. fig. 6).

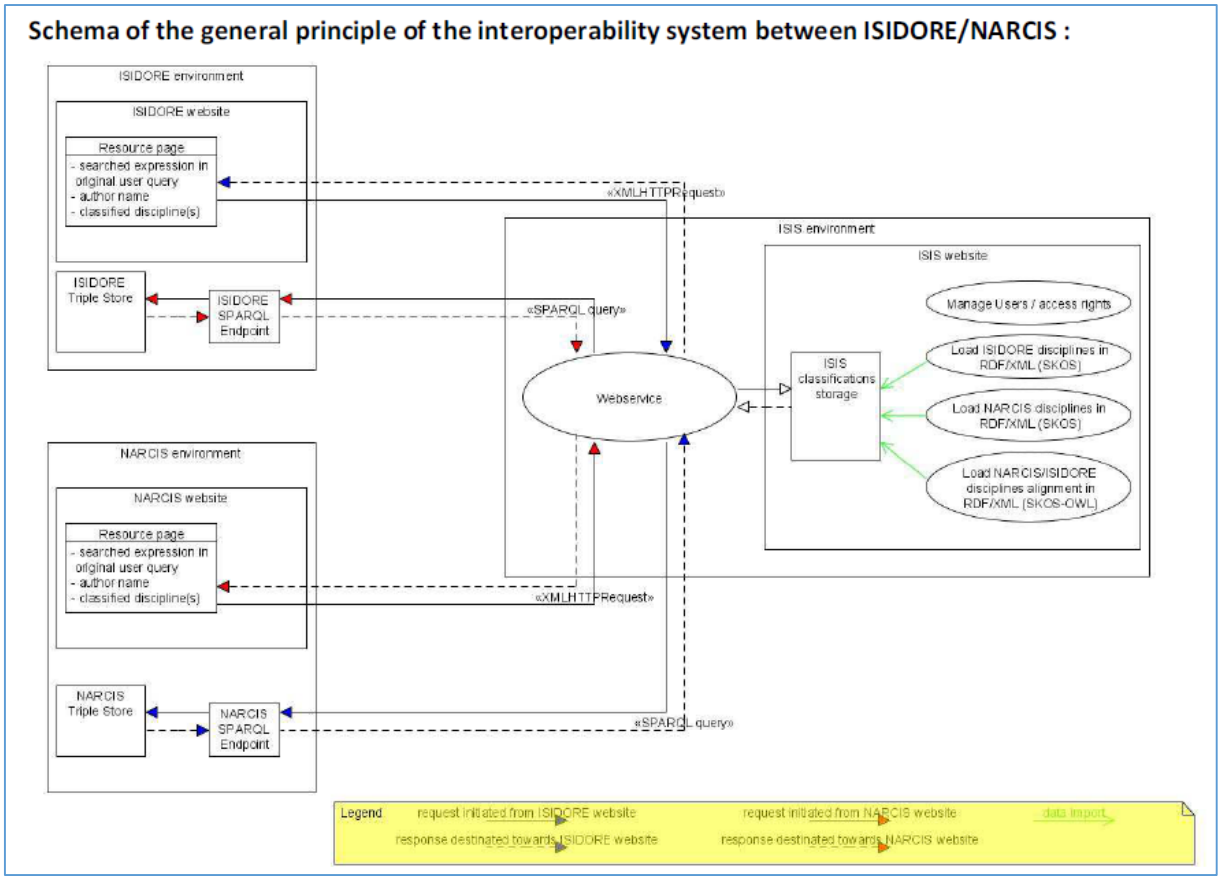


Figure 6 : Interoperability between ISIODRE (Huma-Num) and NARCIS (DANS)

This proof of concept demonstrates the compatibility between ISIDORE and OPENAIRE. In other words, a query in OPENAIRE can dynamically search in ISIDORE and/or in other SparqlEndPoints managed by other stakeholders. This kind of decentralized architecture based on several SparqlEndpoints is more resistant to failure and ensures scalability and sustainability.

3.6. Volume

3.6.1. Resources

ISIDORE was opened in December 2010 with 37 collections and 800 000 resources. In January 2016, 96 collections from 3559 sources provide open access to 4 000 000 documents (scientific papers, multimodal data, maps, PhD dissertations, archives, etc.).

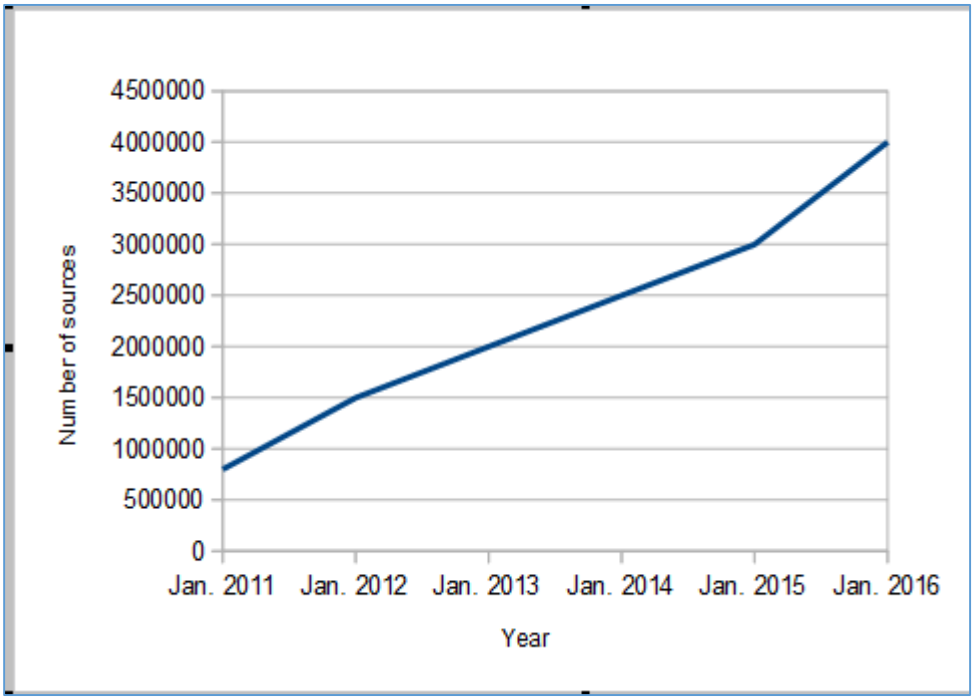


Figure 7 : Resource growth

More than 300 languages are represented; the ten main languages are shown in table 3. The full list of languages can be obtained by the following SPARQL query at this URL: http://www.rechercheisidore.fr/sparql

```
SELECT DISTINCT ?lang count(DISTINCT ?s) as ?nbdoc
{ ?s <http://purl.org/dc/terms/language> ?lang.
FILTER(regex(?lang, " iso639-3"))
}
ORDER BY DESC(?nbdoc)
```

Table 3: Ten main languages in ISIDORE

LANGUAGES	NUMBER of DOCUMENTS
French	3300171
English	210483
Spanish	123291
German	53399
Latin	47732
Italian	29814
Arabic	9818
Persian	8237
Portuguese	7469
Dutch	6389

3.6.2. Consultations

For the year 2015, the total number of consultations was around 1 200 000. Users come from 23 countries and the five main countries are, in descending order, France, Belgium, Canada, the USA, and Italy.

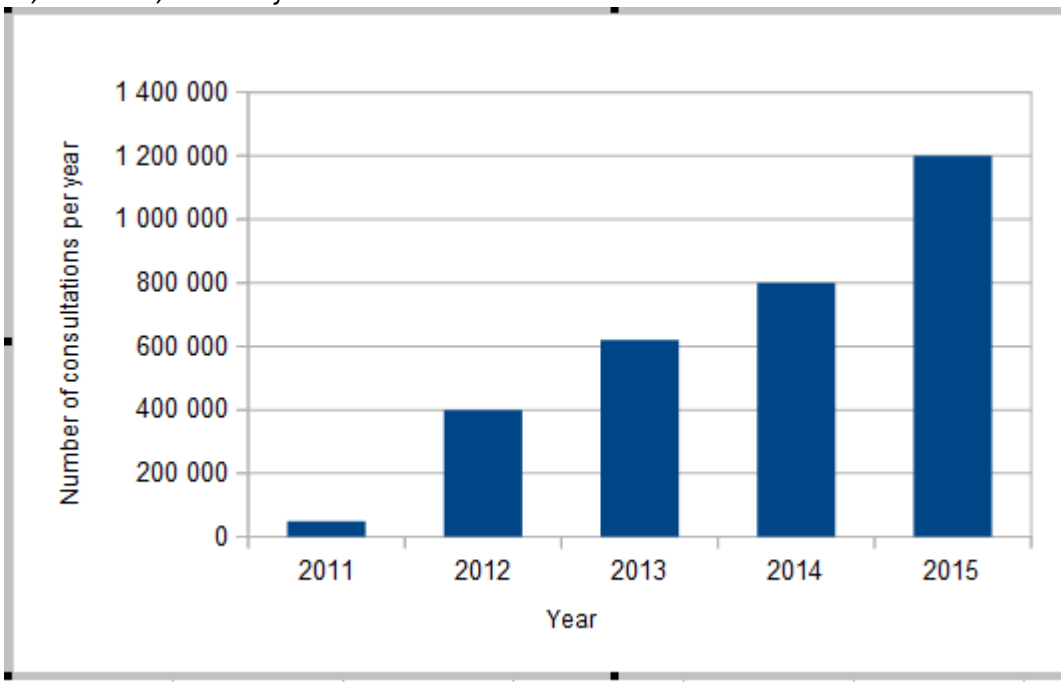


Figure 8: Annual consultations

3.7. ISIDORE in a Nutshell

- **Targeted harvesting** of metadata and scientific data structured according to international and available standards in open access ;
- **Standardization** of metadata and enrichment of data relying on vocabularies recognized in the community;
- **Multilingual (English, French, Spanish) search GUI** exploiting the richness of structured data and vocabularies to make the user an actor of his search;
- **SparqlEndPoint** on sources and indexed data;
- **OPENAIRE compatibility;**
- **Sustainability;**
- **API and widgets**.

3.8. ISIDORE Vocabulary

Table 2: Vocabularies used in ISIDORE

VOCABULARIES	PROPERTIES	COMMENTS
Dublin Core (dcterms)	*	All properties
DC qualified(dcmi)	*	All properties
FOAF	Agent	

	familyName	
	givenName	
	homepage	
	isprimaryTopicOf	
	logo	
	mbox	
	name	
	primaryTopicOf	
	sha1	
	title	
	TopicOf	
GEONAMES	alternateName	
	childrenFeatures	
	countryCode	
	featureClass	
	featureCode	
	locationMap	
	name	
	nearbyFeatures	
	neighbouringFeatures	
	officialName	
	parentADM1	
	parentADM2	
	parentADM3	
	parentADM4	
	parentCountry	
	parentFeature	
	population	
	postalCode	
	shortName	
	wikipediaArticle	
ORE	aggregates	
	isAggregatedBy	
	similarTo	
SKOS	altLabel	
	broader	
	broadMatch	
	changeNote	
	closeMatch	
	definition	
	editorialNote	

	exactMatch	
	example	
	hasTopConcept	
	historyNote	
	inScheme	
	member	
	narrower	
	notation	
	note	
	prefLabel	
	related	
	scopeNote	
RDFS	label	

4. General Synthesis

The French open access research data ecosystem can be represented as a multi-layer system which aims to deal with the various stages in the scholarly content life-cycle as well as to further a culture of data sharing in the Humanities and Social Sciences.

The first layer offers, through OpenEdition, open journals, open books, blogs in which scholars can submit their scientific papers, conference CFPs or write blog posts. OpenEdition is run by the Centre for open electronic publishing (Cléo), a unit that brings together the Centre National de la Recherche Scientifique (CNRS), the Université d'Aix-Marseille, the École des Hautes Études en Sciences Sociales (EHESS) and the Université d'Avignon et des Pays de Vaucluse.

Open Archives (HAL, HAL-SHS, TEL, Theses) constitute the second layer in which scholars deposit their scientific papers published by journals managed by OpenEdition or other publishers. CCSD and ABES (Bibliographic Agency of Higher Education) run these services.

The third layer is dedicated to managing and securing scientific metadata and data with NAKALA. These metadata and data are the product of descriptions or experiments which were presented by scholars in scientific publications.

The fourth layer is constituted by the ISIDORE and NAKALONA platforms which tag, enrich and push data and metadata produced by the other three layers. Both platforms share these data by providing access to two SparqlEndpoints, making them interoperable with other platforms such as NARCIS, OPENAIRE or any platforms based on RDF triplestore. It must be pointed out that ISIDORE harvests far beyond French data providers since American, Belgian, Spanish, and Italian data providers (library, research centers, etc.) have requested to be harvested. The Very Large Facility Huma-Num runs the third and fourth layers of services.