A new textual search engine to discover VizieR catalogs

CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

Textual search is a part of the VizieR indexation which completes the position indexation and the keyword search resulting from SQL queries. This new capability extends the capacity of the current engine with a natural language approach like in Google, or bumblebee (ADS). The new version – still in alpha – uses the Elasticsearch engine, an Open Source Software that indexes documents with a grammar and a textual search analyser.

The query supported is a NO-SQL language including strict or fuzzy search and available through an HTTP RESTful API. Then, a fine configuration adapted to the different data (authors, DOI, abstract, date...) is needed to improve the indexation.

Resources indexed are the ReadMes which describe the VizieR catalogs. A ReadMe is a structured ASCII file containing the basis metadata : authors, title, keywords, abstracts but also tables and column descriptions. We explain the new implementation from the data origin to the final users.

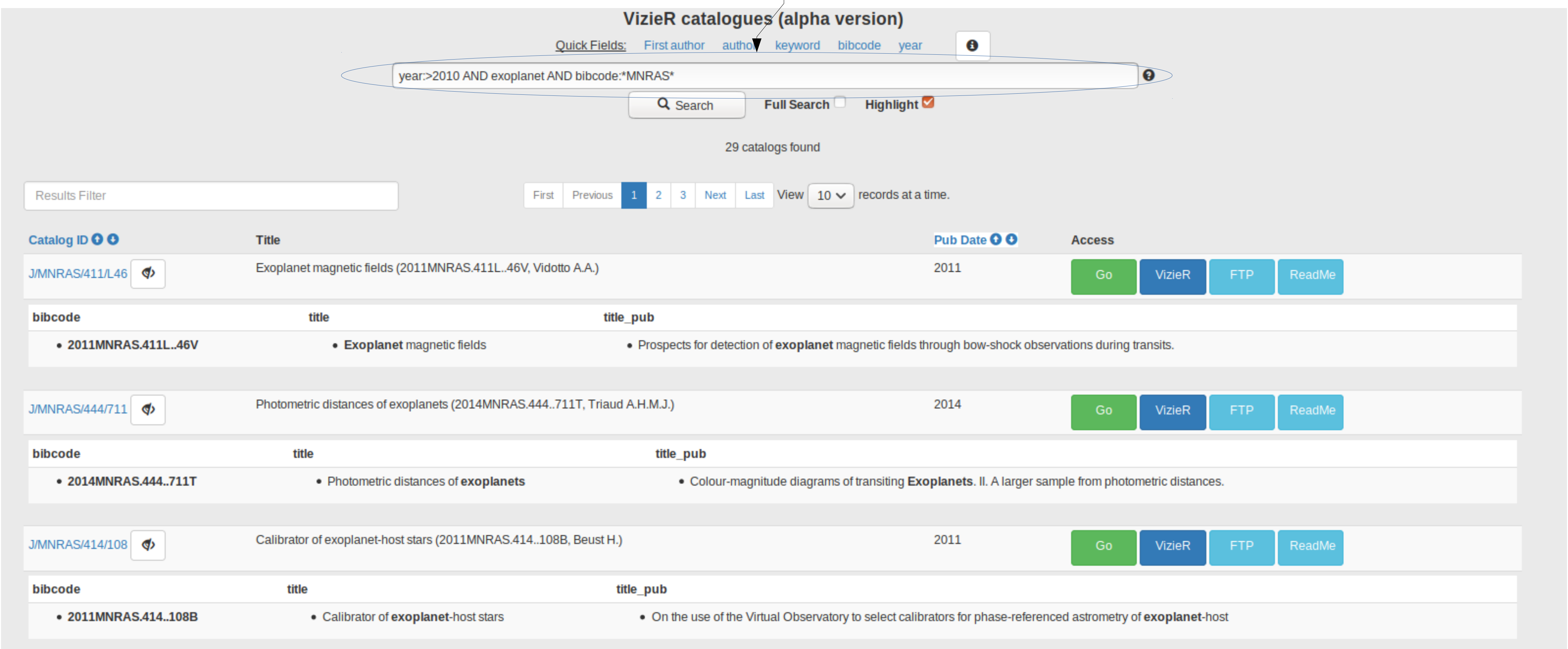
VizieR will use Elasticsearch for textual search
in addition to the other searches like search by position

elasticsearch

A word on Elasticsearch operation

Elasticsearch, based on Apache Lucene engine is allowing to implement a flexible search engine easily. In this purpose there are three phases to follow : index configuration, data ingestion and finally search.
The first part is to configure an index by choosing the way data should look like and how they will be treated. Once done, you need to format your data into JSON, the data format for Elasticsearch communication. Then you are ready to ingest them in Elasticsearch engine. Finally you can use the Elasticsearch API to query the index.

Fig.1: A new interface
Search for Year:>2010 AND exoplanet AND bibcode:*MNRAS*



VizieR catalogues (alpha version)

Quick Fields: First author author keyword bibcode year

year:>2010 AND exoplanet AND bibcode:*MNRAS*

Search Full Search □ Highlight ☑

29 catalogs found

Results Filter

First Previous 1 2 3 Next Last View 10 records at a time.

Catalog ID	Title	Pub Date	Access
J/MNRAS/411/L46	Exoplanet magnetic fields (2011MNRAS.411L..46V, Vidotto A.A.)	2011	Go VizieR FTP ReadMe
J/MNRAS/444/711	Photometric distances of exoplanets (2014MNRAS.444..711T, Triaud A.H.M.J.)	2014	Go VizieR FTP ReadMe
J/MNRAS/414/108	Calibrator of exoplanet-host stars (2011MNRAS.414..108B, Beust H.)	2011	Go VizieR FTP ReadMe

bibcode	title	title_pub
• **2011MNRAS.411L..46V**	• **Exoplanet** magnetic fields	• Prospects for detection of **exoplanet** magnetic fields through bow-shock observations during transits.

bibcode	title	title_pub
• **2014MNRAS.444..711T**	• Photometric distances of **exoplanets**	• Colour-magnitude diagrams of transiting **Exoplanets**. II. A larger sample from photometric distances.

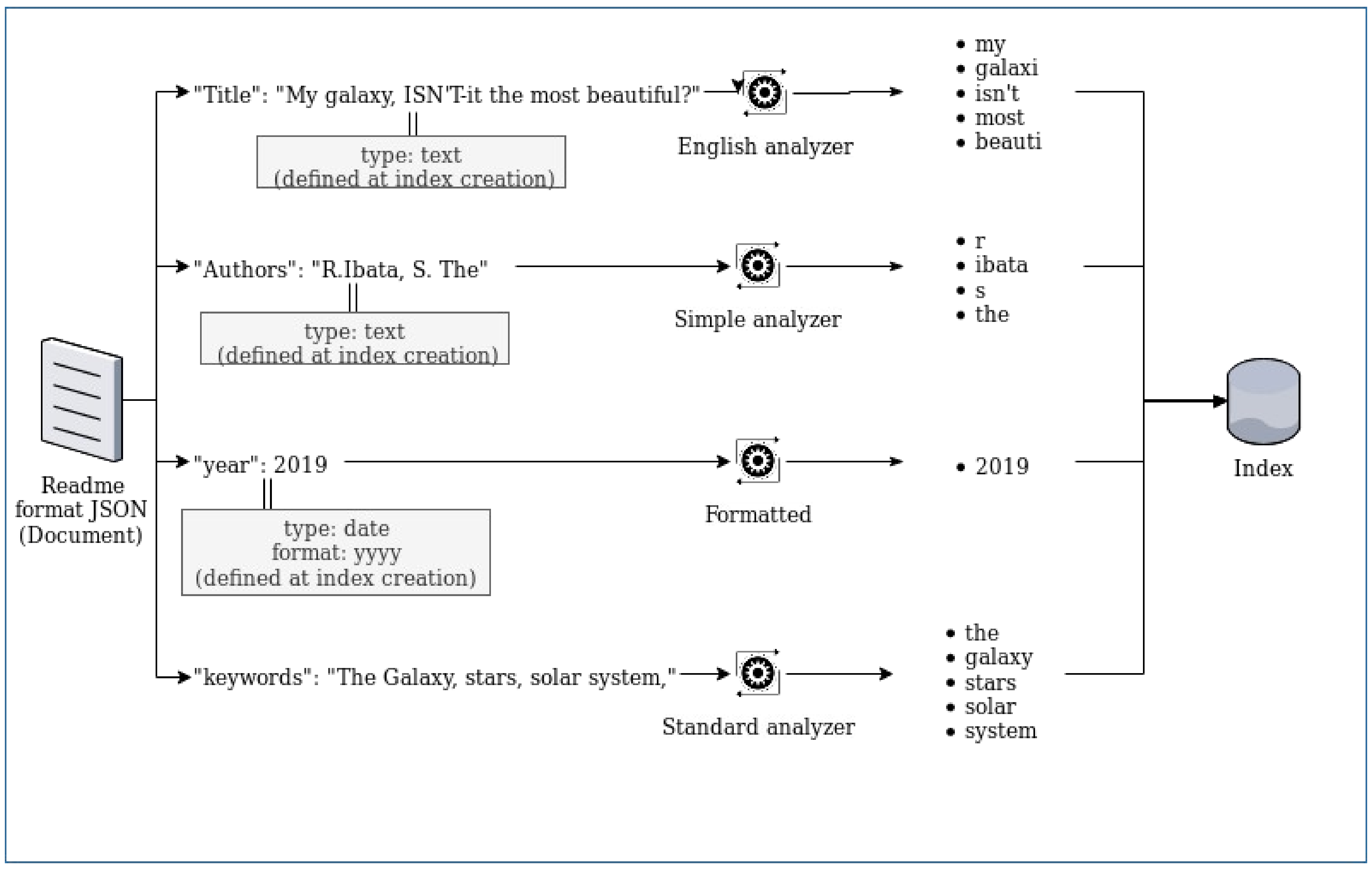
bibcode	title	title_pub
• **2011MNRAS.414..108B**	• Calibrator of **exoplanet**-host stars	• On the use of the Virtual Observatory to select calibrators for phase-referenced astrometry of **exoplanet**-host

A new alpha interface to access the « VizieR bazar »

The fig.1 on the right shows an example of Elasticsearch usage. It gives you access to a complete grammar for requests, allowing to use operators, fuzziness (levenshtein distance), phrase analysis, proximity (number of words between words of a phrase) and other. All that you can combine the way you want.

The chosen vocabulary is based on ADS field naming : author, title...
Now it is also possible to search on doi, orcid, year of publication.

The alpha version is available here :
http://cdsarc.u-strasbg.fr/viz-bin/cat2



"Title": "My galaxy, ISN'T-it the most beautiful?"
type: text (defined at index creation)
English analyzer
- my
- galaxi
- isn't
- most
- beauti

"Authors": "R.Ibata, S. The"
type: text (defined at index creation)
Simple analyzer
- r
- ibata
- s
- the

Readme format JSON (Document)

"year": 2019
type: date format: yyyy (defined at index creation)
Formatted
- 2019

Index

"keywords": "The Galaxy, stars, solar system,"
Standard analyzer
- the
- galaxy
- stars
- solar
- system

fig.2: ElasticSearch documents analysis (indexation)

Analysis

On fig.2 you can look at an example of a document analysis, here a Readme. It shows 4 fields treated differently for the indexation and the effective outcome that will be indexed.

Current work

Elasticsearch ingestion and search are ready for VizieR in local CDS architecture and will be proposed to the outside (VizieR mirrors and other workflows).
Currently, the mirroring architecture has been tested locally in a docker environment.

Perspectives

In the future, Elasticsearch for VizieR may be used to populate the VizieR catalogs in the VO registry. Some tests are planned to use Elasticsearch to feed the chatbot experimented to access the CDS servers.

ADASS2019, October 6-10, Groningen, the Netherlands – Observatoire astronomique de Strasbourg

Bisch.Y, Landais.G, Schaaff.A
yann.bisch@astro.unistra.fr

Ack : Thanks to all people who contributed to design the ElasticsearchVizieR search engine

VizieR

cnrs dépasser les frontières

UNIVERSITÉ DE STRASBOURG