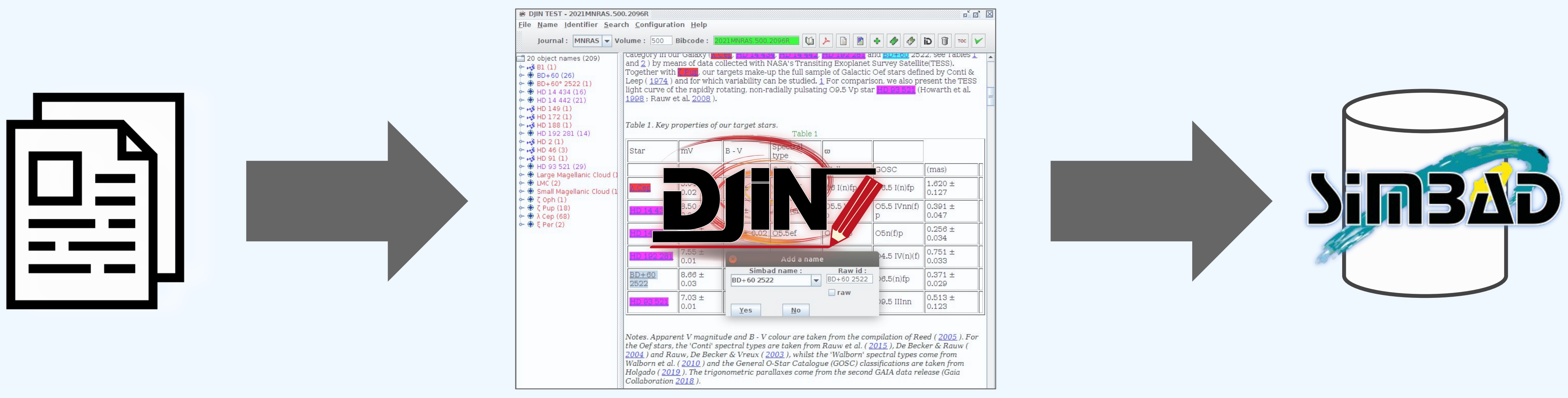
CDS homogenisation of metadata from publishers

Centre de Données astronomiques de Strasbourg

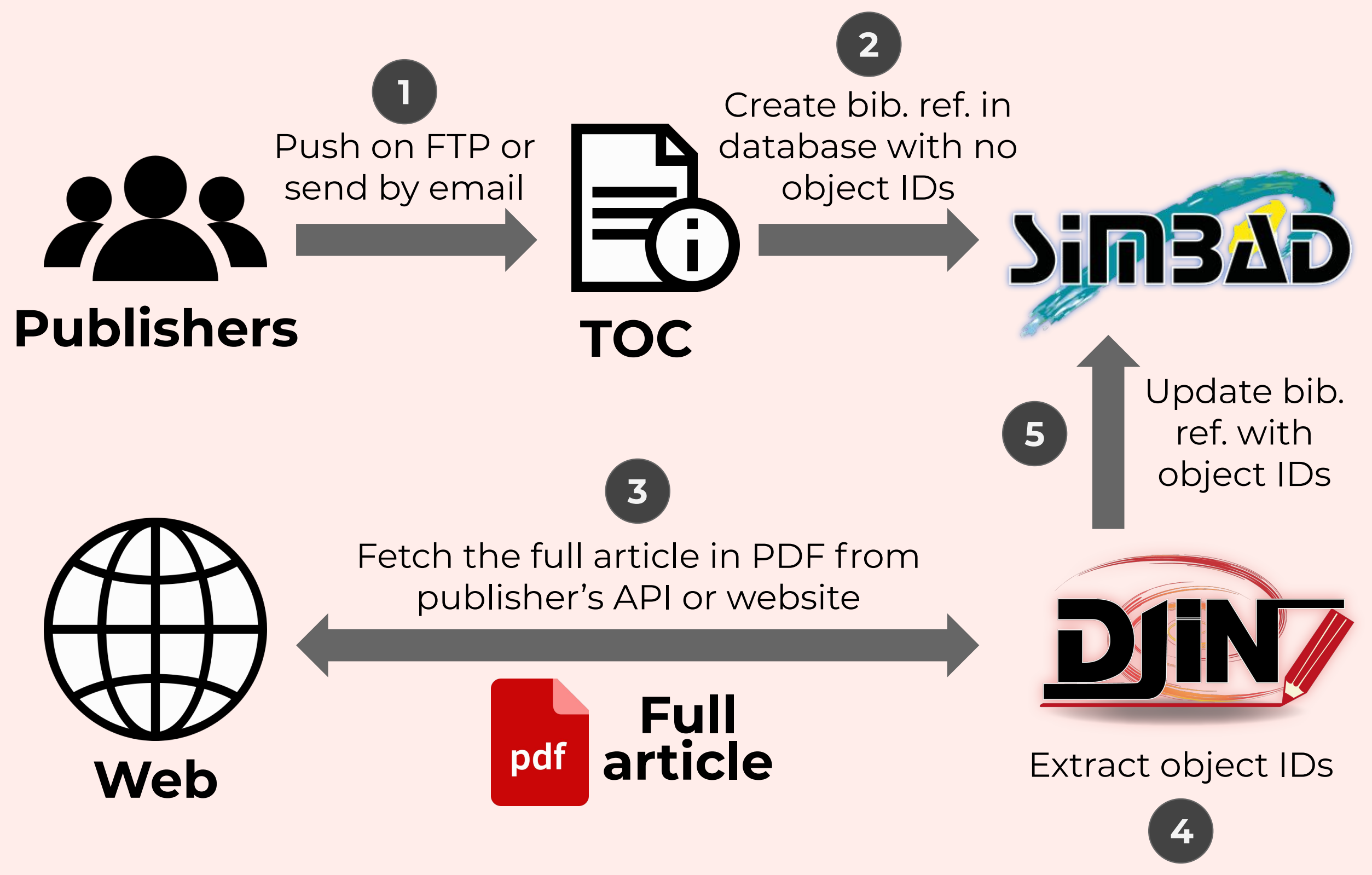
Grégory Mantelet
gregory.mantelet@astro.unistra.fr
Oberto A., Neuville M., Lesteven S., Allen M.

What purpose?

- **Articles analysed**…
- …for astronomical **object names extraction**
- …by **DJIN** (a semi-automatic software)
- …to update the objects stored in the **SIMBAD database**
- …with new identifiers, positions
- …and **links between objects and their bibliographic references**.



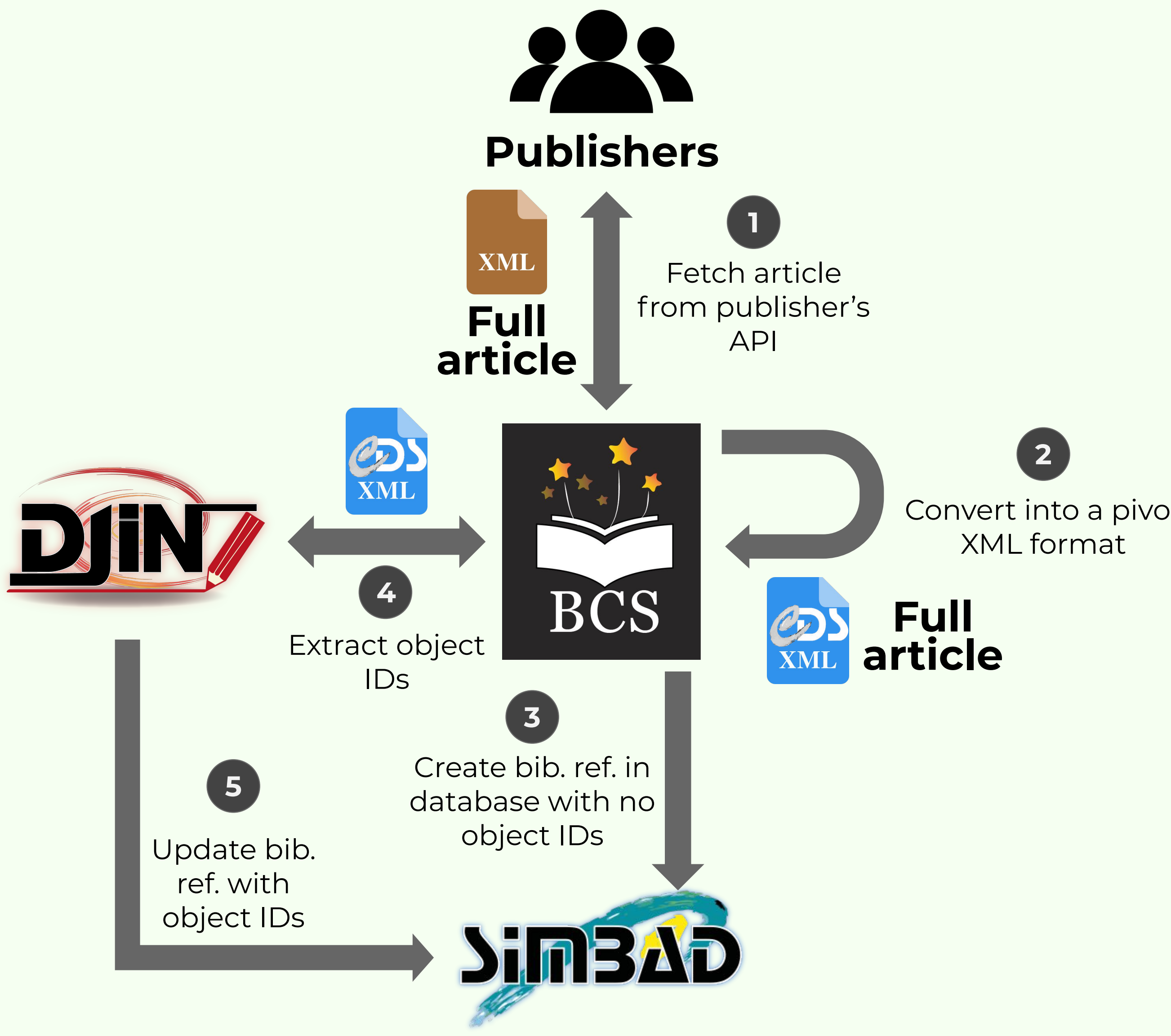
Now


1
Push on FTP or send by email
Publishers
TOC
2
Create bib. ref. in database with no object IDs
SIMBAD
5
Update bib. ref. with object IDs
3
Fetch the full article in PDF from publisher's API or website
Web
pdf
Full article
DJIN
Extract object IDs
4

What change?

	Now	In progress
Format	**PDF** • not reliable for text extraction but for display and print, • different from one publisher to another, • very sensitive to change	**XML** • easy text extraction (XPath) • easy conversion into other formats (XSLT) • if following a **standard**, XPath+XSLT reusable (JATS)
Files	**2 files** with a different way to get each of them depending on the publisher	**Only 1 file** with a different way to get it depending on the publisher
Workflow	**Completely manual + Scripts** (Bash/Python) get and process the TOC + get the PDF article from the Web	**Semi-automatic + Internal Web Site** fetch+convert on demand with the BCS when available
Get	**TOC: Push** (pushed by publishers on our FTP or by email) **PDF: Push** on our FTP or **PDF: Pull** from web site	**Pull** (get XML from publishers) *As much as possible, but not yet possible with most publishers*
Reusable?	No	**Yes** *e.g. for table extraction (VizieR)*

In progress

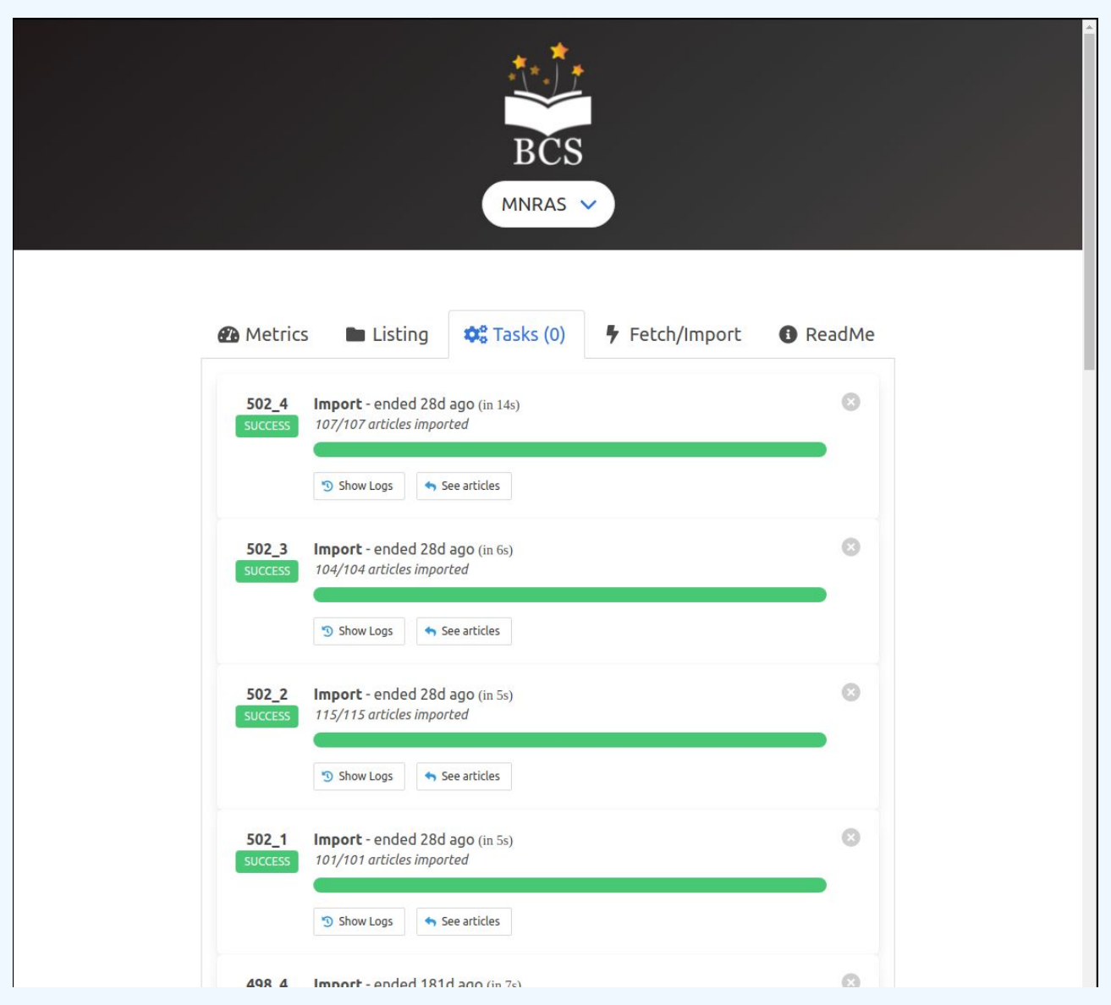

Publishers
XML
Full article
1
Fetch article from publisher's API
2
Convert into a pivot XML format
XML
Full article
BCS
XML
DJIN
4
Extract object IDs
3
Create bib. ref. in database with no object IDs
5
Update bib. ref. with object IDs
SIMBAD

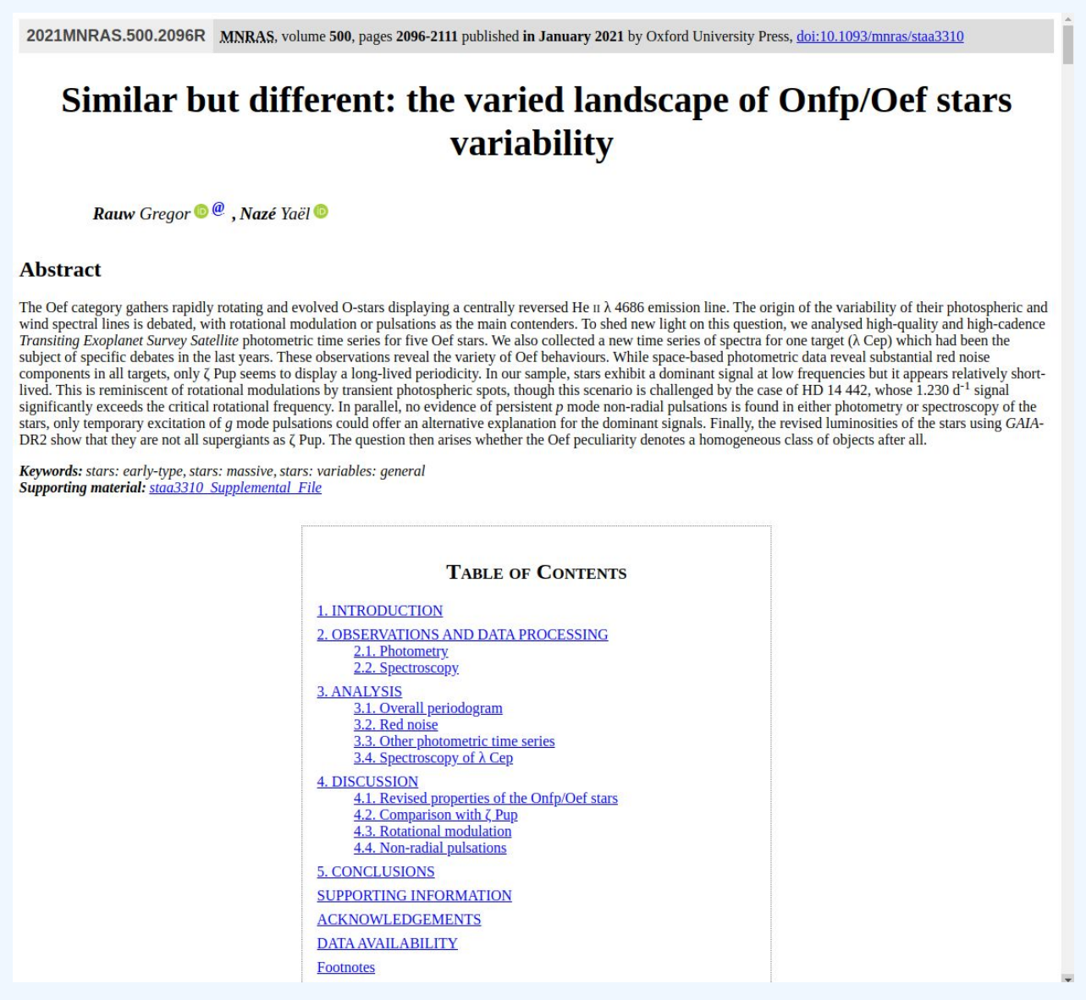
But…

- **Not all** journals provide their articles **in XML**
- If existing, the **XML** is **not always accessible for us**
- If existing, **often not standard** (though there are standards like **JATS**)
- **Missing** Pull **Service or API** to fetch XML articles
- **Missing notifications** about **complete** new volumes/issues

BCS

Bibliographic Center Supervisor
(Internal CDS Web Site)




Similar but different: the varied landscape of Onfp/Oef stars variability
Abstract
Table of Contents

LISA IX Library and Information Services in Astronomy

Observatoire astronomique de Strasbourg

cnrs

Université de Strasbourg

CDS homogenisation of metadata from publishers – Royal Astronomy Society – 14-18 June 2021 - CDS – Observatoire astronomique de Strasbourg