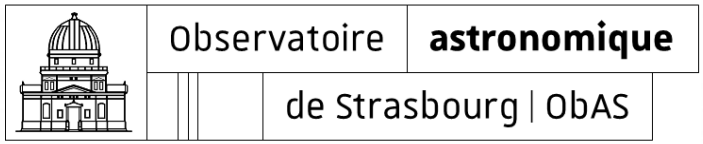
DJIN: de XML à XCDS

Grégory Mantelet[1]

[1]CDS (Centre de Données astronomiques de Strasbourg)

9 Mai 2019



XML & Co - 1/3

XML *(e**X**tensible **M**arkup **L**anguage)*

Méta-language de balisage générique.

Il est dit *extensible* car il permet de définir sa propre structure, ses propres balises. . . et donc, son propre vocabulaire.

XML & Co - 2/3

Ce vocabulaire doit être défini par un schema. 2 formats possibles:

- ***DTD** (**D**ocument **T**ype **D**escription)*

  Description simple de la structure, des balises et attributs autorisés dans un document XML.

- ***XSD** (**X**ML **S**chema **D**escription)*

  Même chose qu'une DTD mais en plus précis et plus sophistiqué.

XML & Co - 3/3

Il est possible de *styliser* ou plutôt *transformer* un document XML en n'importe quel autre format textuel (e.g. texte, CSV, XML, HTML) avec le language XML suivant:

- ***XSL*** *(e**X**tensible **S**tylesheet **L**anguage)*

  Une feuille de style XSL est un fichier qui décrit comment doivent être transformés des documents XML qui suivent le même schéma.

  Cela regroupe d'autres notions liées au language XML:

  - ***XPath*** *(XML Path ; language de navigation dans un document XML)*
  - ***XSLT*** *(XSL Transforms ; language de transformation)*

XCDS - Structure

XSCDS = ***X****ML* ***CDS*** *dédié à la biblio*

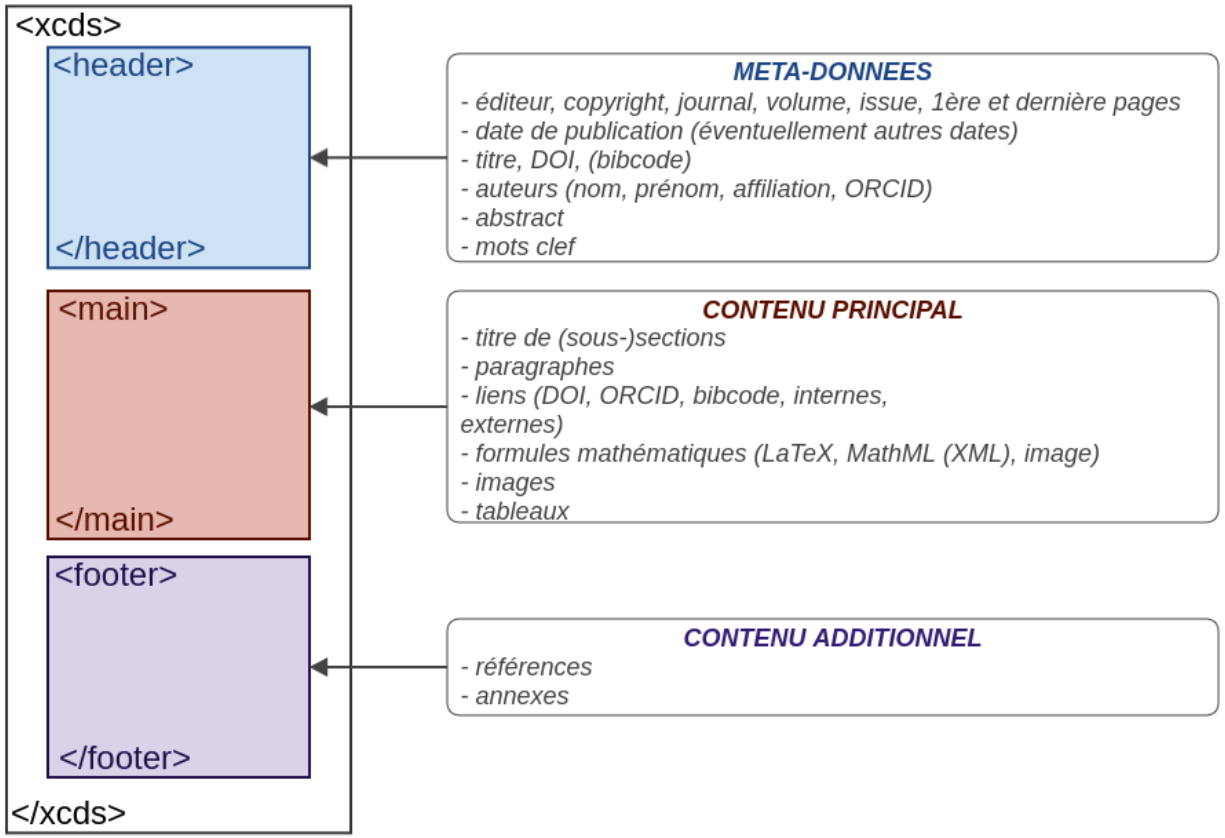


Figure 1: Structure du XCDS

XCDS - Génération

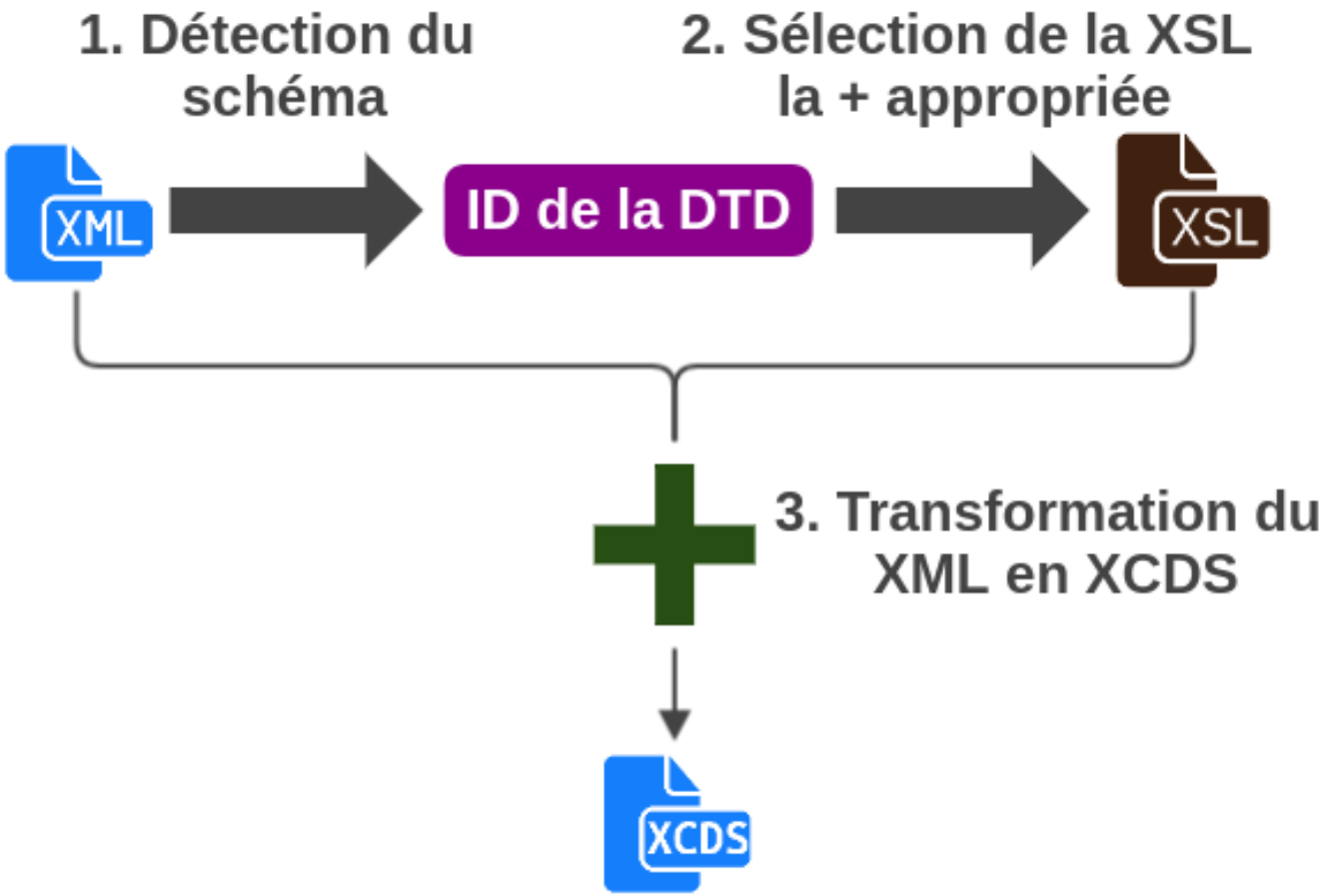


Figure 2: Conversion d'un XML en XCDS

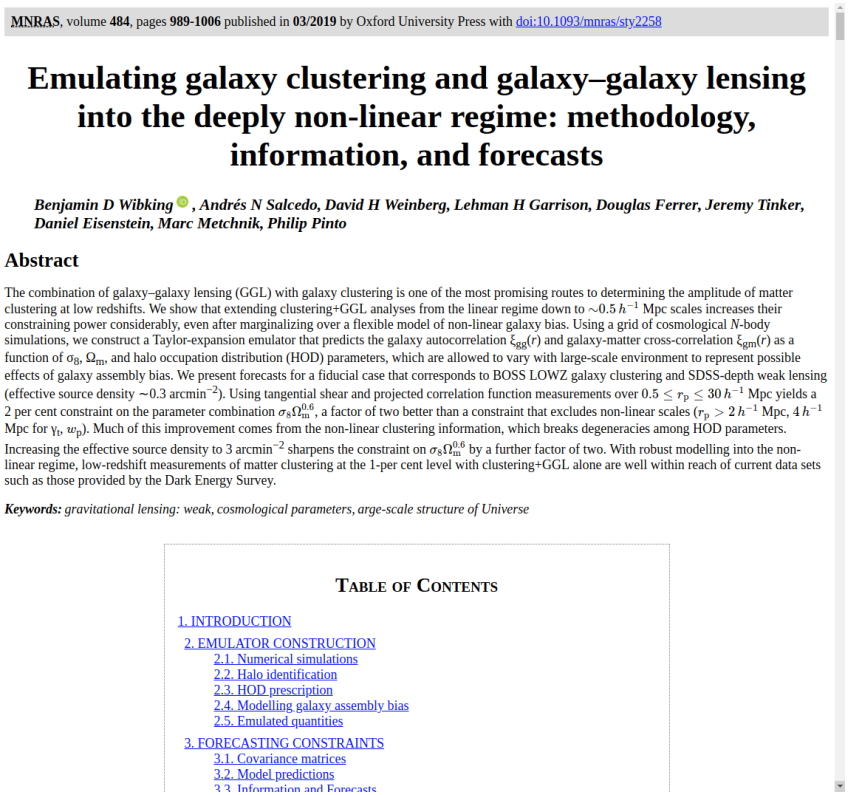
XCDS - Vérification

Une fois un XCDS généré, le résultat devra être validé afin de s'assurer que des outils comme Djin arrivent à bien le lire.

- *Encore à faire (quand le format XCDS sera suffisamment stable):*
  - écrire la XSD/DTD de notre format XCDS

XCDS - Usage - 1/2

- Visualisation dans:
  - DJIN *(cf démo d'Anaïs)*
  - un navigateur Web



MNRAS, volume **484**, pages **989-1006** published in **03/2019** by Oxford University Press with doi:10.1093/mnras/sty2258

Emulating galaxy clustering and galaxy–galaxy lensing into the deeply non-linear regime: methodology, information, and forecasts

Benjamin D Wibking*, *Andrés N Salcedo, David H Weinberg, Lehman H Garrison, Douglas Ferrer, Jeremy Tinker, Daniel Eisenstein, Marc Metchnik, Philip Pinto

Abstract

The combination of galaxy–galaxy lensing (GGL) with galaxy clustering is one of the most promising routes to determining the amplitude of matter clustering at low redshifts. We show that extending clustering+GGL analyses from the linear regime down to $\sim 0.5\,h^{-1}$ Mpc scales increases their constraining power considerably, even after marginalizing over a flexible model of non-linear galaxy bias. Using a grid of cosmological *N*-body simulations, we construct a Taylor-expansion emulator that predicts the galaxy autocorrelation $\xi_{gg}(r)$ and galaxy-matter cross-correlation $\xi_{gm}(r)$ as a function of σ_8, Ω_m, and halo occupation distribution (HOD) parameters, which are allowed to vary with large-scale environment to represent possible effects of galaxy assembly bias. We present forecasts for a fiducial case that corresponds to BOSS LOWZ galaxy clustering and SDSS-depth weak lensing (effective source density ~ 0.3 arcmin^{-2}). Using tangential shear and projected correlation function measurements over $0.5 \leq r_p \leq 30\,h^{-1}$ Mpc yields a 2 per cent constraint on the parameter combination $\sigma_8\Omega_m^{0.6}$, a factor of two better than a constraint that excludes non-linear scales ($r_p > 2\,h^{-1}$ Mpc, $4\,h^{-1}$ Mpc for γ_t, w_p). Much of this improvement comes from the non-linear clustering information, which breaks degeneracies among HOD parameters. Increasing the effective source density to 3 arcmin^{-2} sharpens the constraint on $\sigma_8\Omega_m^{0.6}$ by a further factor of two. With robust modelling into the non-linear regime, low-redshift measurements of matter clustering at the 1-per cent level with clustering+GGL alone are well within reach of current data sets such as those provided by the Dark Energy Survey.

Keywords: *gravitational lensing: weak, cosmological parameters, arge-scale structure of Universe*

Table of Contents

Figure 3: Visualisation dans un navigateur Web

XCDS - Usage - 2/2

- Extraction aisée de:
  - table des matières *(en Parfile ou autre format texte)*
  - méta-données attachées à l'article (e.g. DOI, nom du journal, auteurs, . . .)
  - tables
  - . . .

Centraliser les journaux - 1/2

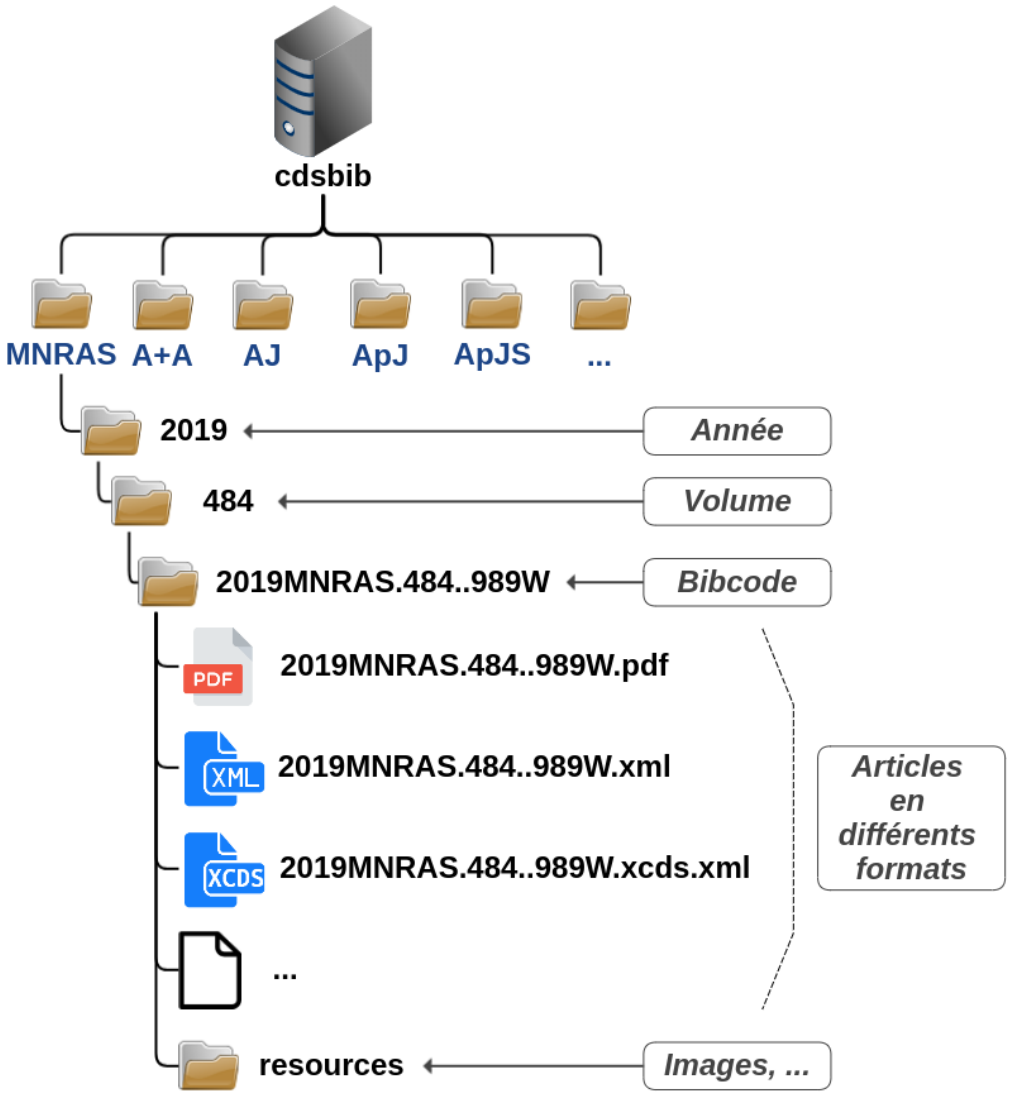


Figure 4: Hiérarchie proposée pour les journaux

Centraliser les journaux - 2/2

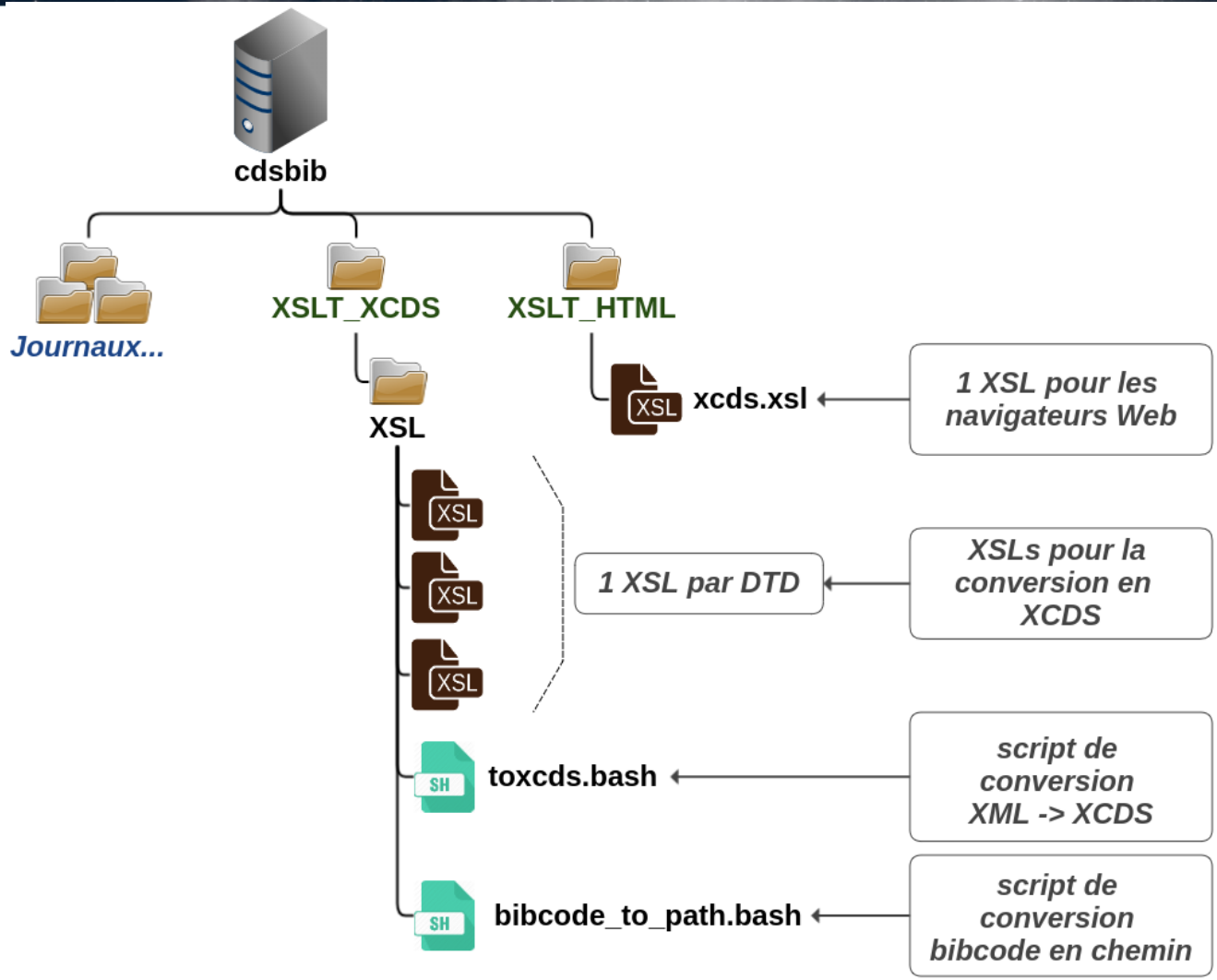


Figure 5: Scripts et XSL

Assurer la pérénité - 1/2

- Négociation avec les éditeurs
  - s'assurer que les formats ne changent pas trop régulièrement
  - quand ils changent de façon permanente, nous avertir
- Se baser sur l'identification des schémas XML des articles (i.e. DTD ou XSD)

  Exemple: *A&A, AJ & Co et MNRAS suivent chacun une DTD particulière. Une fois l'ID de la DTD trouvé, il nous suffit de sélectionner dans notre liste de XSLs celle qui correspond.*

Assurer la pérénité - 2/2

La DTD de A&A, d'AJ & Co et de MNRAS sont très similaires car elles se basent toutes sur un ensemble de balises. . . .

- ***JATS*** *(**J**ournal **A**rticle **T**ag **S**uite)*
  - https://jats.nlm.nih.gov/
    > [...] defines a set of XML elements and attributes
    > for tagging journal articles and describes three
    > article models.
  - structure et ensemble de balises bien définis et adaptés pour des publications scientifiques
  - plusieurs éditeurs utilisent maintenant la version officielle (e.g. MNRAS) ou une adaptation d'une version préliminaire (e.g. A&A et IoP)
  - DTD/XSD, XSLs et documentations/tutoriels disponibles sur le site Web
  - a débuté en 2011 (0.4)
  - dernière version: 1.2 (Février 2019)

De PDF à XML

- Encore en recherche d'un outils idéal
  - mais tous les témoignages sur le Web sont unanimes: aucun outils ne fonctionnera complètement et encore plus si on veut l'utiliser pour différentes "versions" de PDF.
- 9 outils testés *(hormis PDFBox déjà utilisé dans DJIN-1 et Grobid dans DJIN-2)*
- Nouvelle version de PDFBox éventuellement à tester
- Sinon Grobid