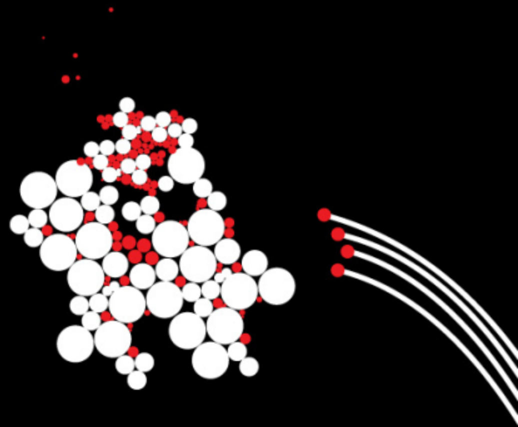
UNIVERSITEIT TWENTE.

COMMIT/

mydatafactory

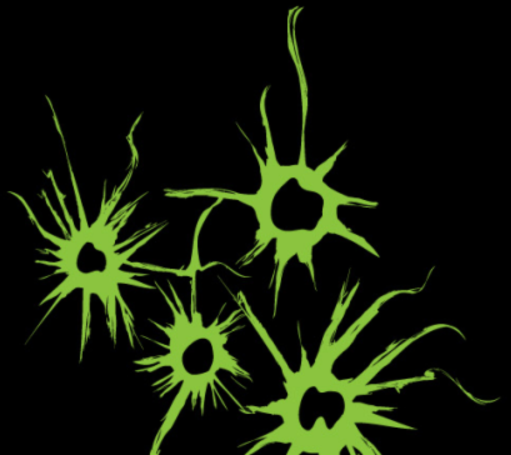


SMART CONSOLIDATION OF PRODUCT INFORMATION

Maurice van Keulen[1], Dolf Trieschnigg[1,2], Brend Wanders[1]

[1]University of Twente, Enschede, Netherlands

[2]Mydatafactory, Meppel, Netherlands

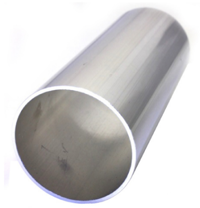


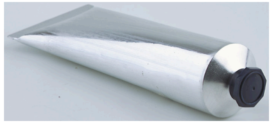
PRODUCT DATA

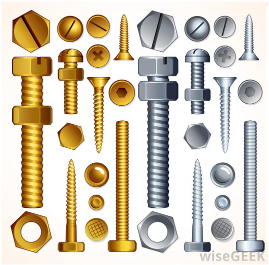
WHAT IS IT AND WHY IS IT A PROBLEM?

What is it

- Data and specification on parts, substances, etc.







Why is it a problem?

- High requirements on data quality
- Errors and duplicates may be costly or even pose health risks

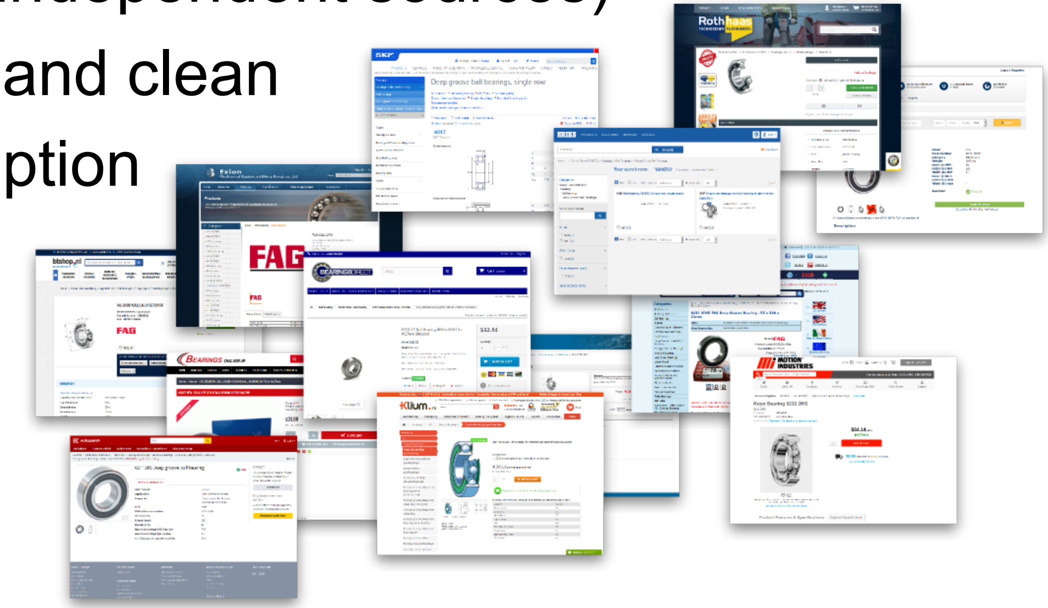
➢ Even so, it is a mess (more on that later!)

PRODUCT INFORMATION CLEANING AND ENRICHMENT

Proposed approach

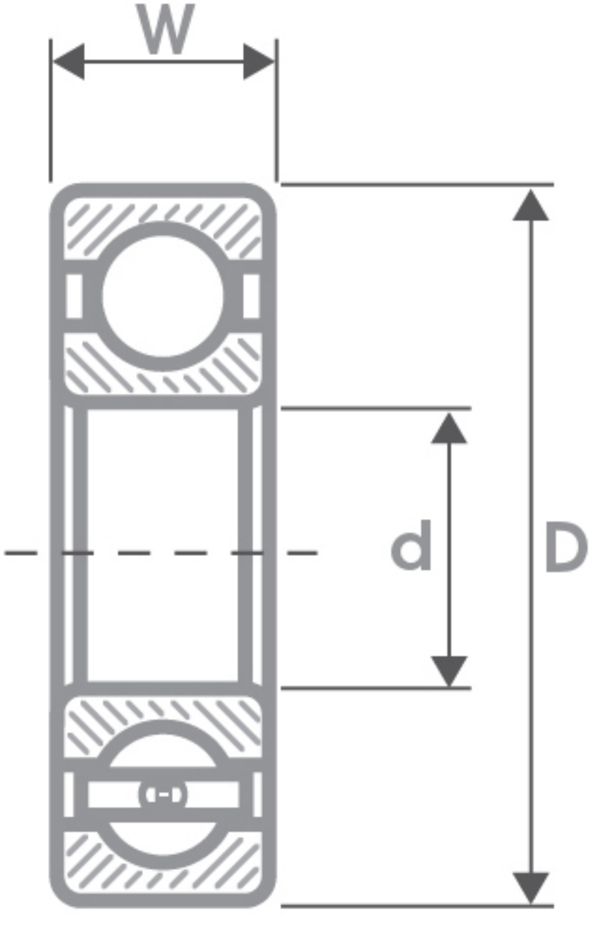
- Given catalogue / database with data on products
- Gather data on the same products from websites (many more or less independent sources)
- Consolidate: merge and clean

➢ One enriched description of the product



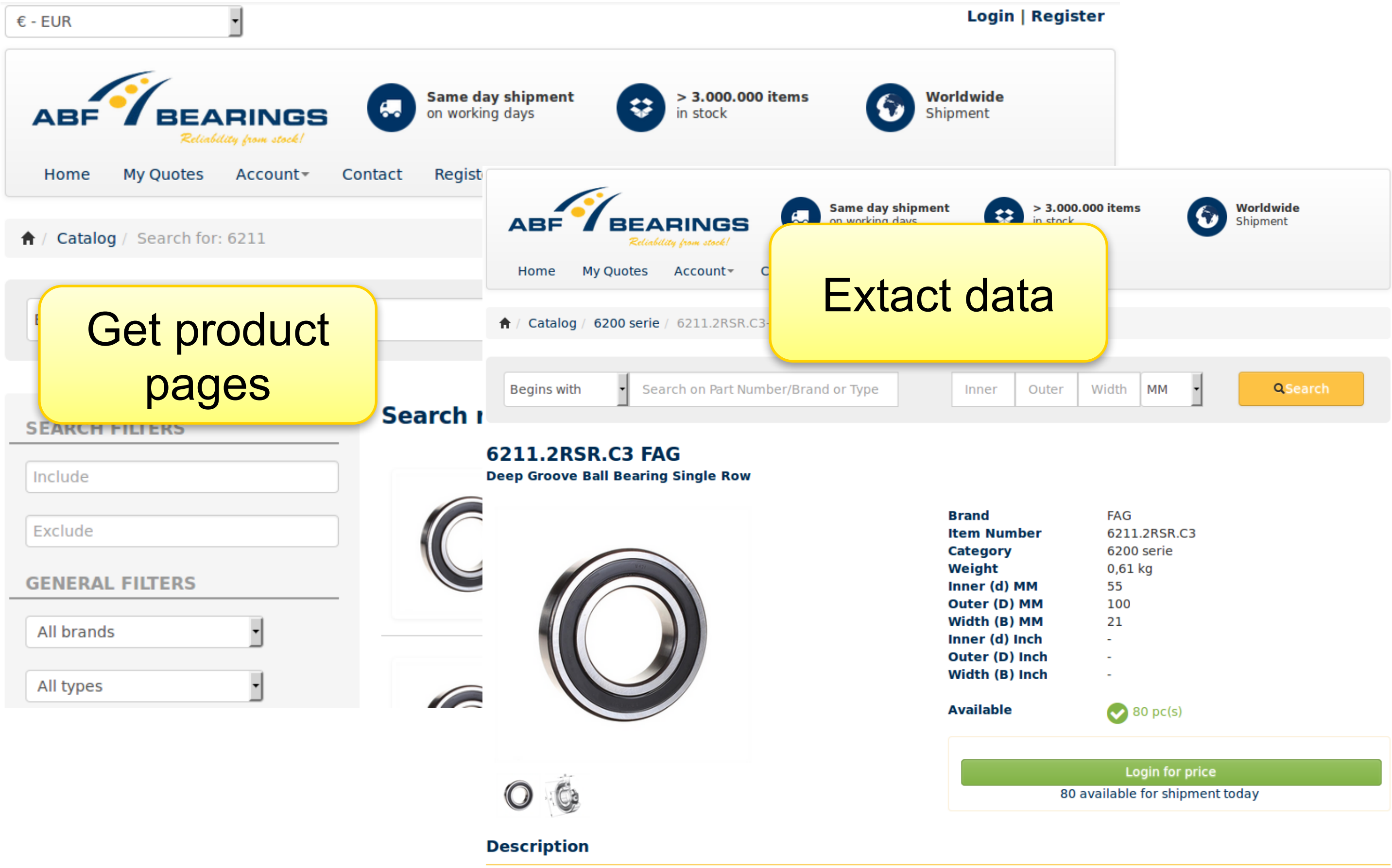
PILOT: BALL BEARINGS

1. GIVEN CATALOGUE / DATABASE WITH DATA ON PRODUCTS



PILOT: BALL BEARINGS

2. GATHER DATA ON THE SAME PRODUCTS FROM WEBSITES; 3. CONSOLIDATE



6211.2RSR.C3 FAG

Deep Groove Ball Bearing Single Row

Brand	FAG
Item Number	6211.2RSR.C3
Category	6200 serie
Weight	0,61 kg
Inner (d) MM	55
Outer (D) MM	100
Width (B) MM	21
Inner (d) Inch	-
Outer (D) Inch	-
Width (B) Inch	-
Available	✓ 80 pc(s)

Login for price

80 available for shipment today

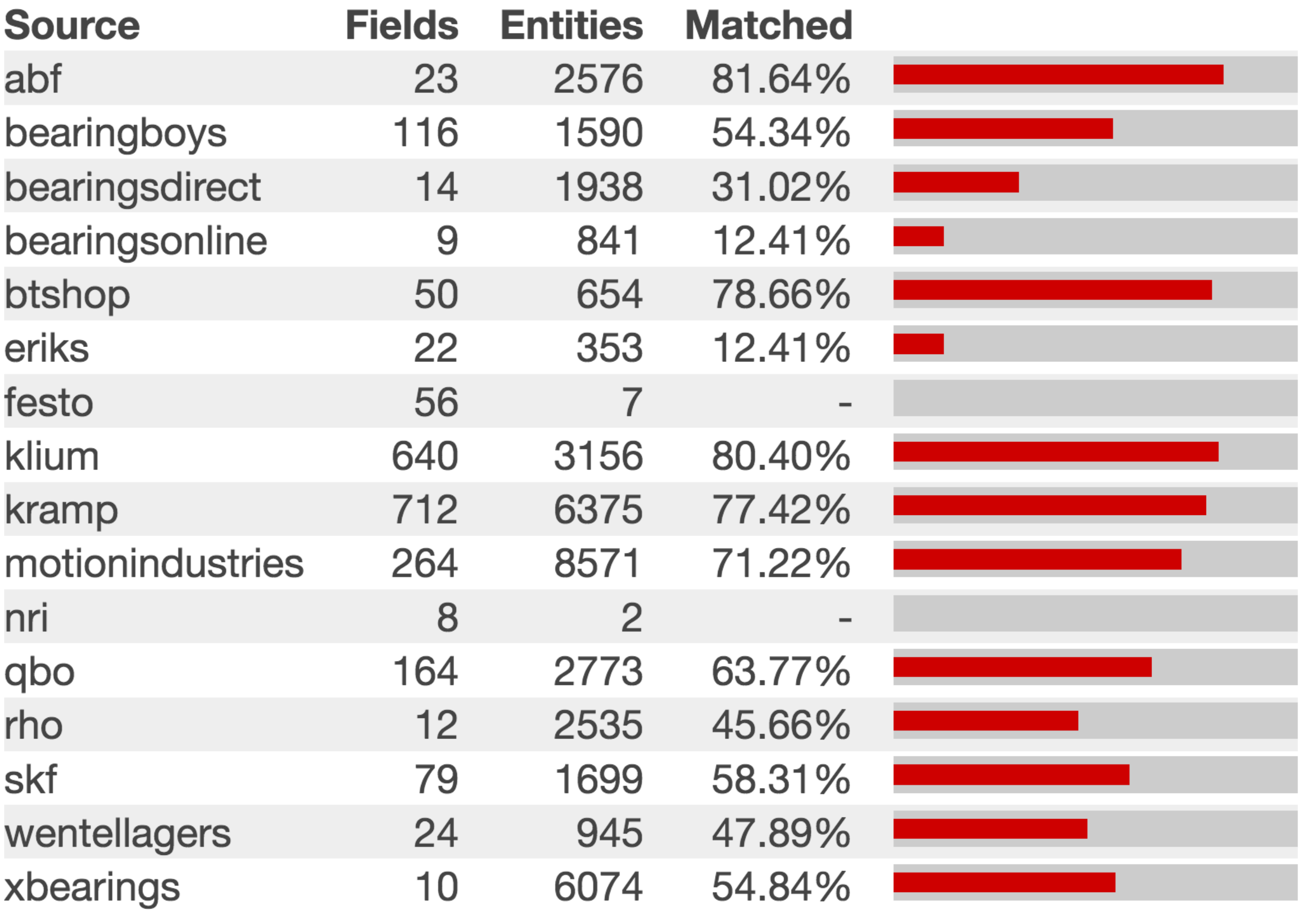
Description

Get product pages

Extact data

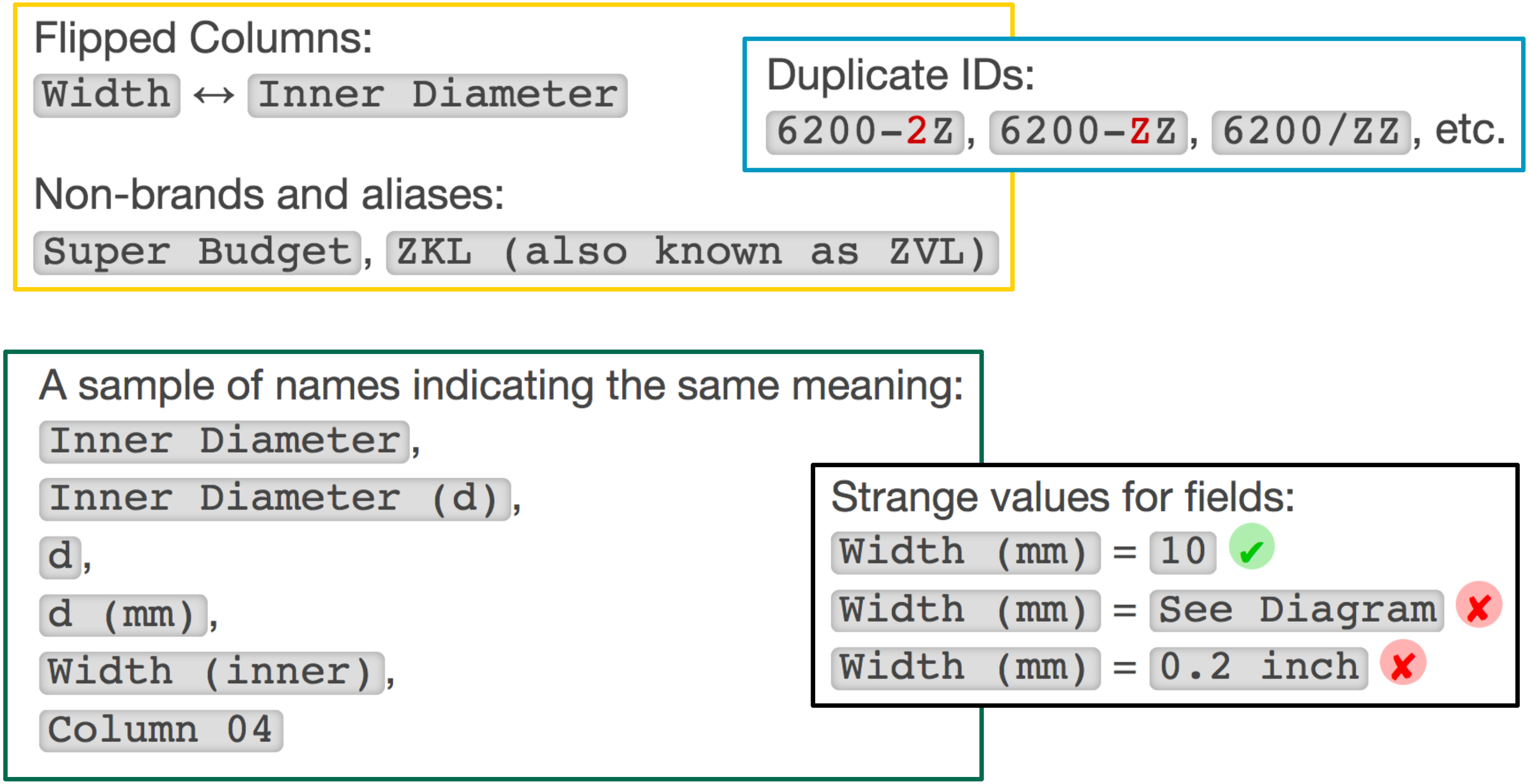
Consolidate (merge, clean)

PILOT: EXPERIENCES



Source	Fields	Entities	Matched
abf	23	2576	81.64%
bearingboys	116	1590	54.34%
bearingsdirect	14	1938	31.02%
bearingsonline	9	841	12.41%
btshop	50	654	78.66%
eriks	22	353	12.41%
festo	56	7	-
klium	640	3156	80.40%
kramp	712	6375	77.42%
motionindustries	264	8571	71.22%
nri	8	2	-
qbo	164	2773	63.77%
rho	12	2535	45.66%
skf	79	1699	58.31%
wentellagers	24	945	47.89%
xbearings	10	6074	54.84%

PILOT EXPERIENCES



Flipped Columns:

`Width` ↔ `Inner Diameter`

Non-brands and aliases:

`Super Budget`, `ZKL (also known as ZVL)`

Duplicate IDs:

`6200-2Z`, `6200-ZZ`, `6200/ZZ`, etc.

A sample of names indicating the same meaning:

`Inner Diameter`,
`Inner Diameter (d)`,
`d`,
`d (mm)`,
`Width (inner)`,
`Column 04`

Strange values for fields:

`Width (mm)` = `10` ✓
`Width (mm)` = `See Diagram` ✗
`Width (mm)` = `0.2 inch` ✗

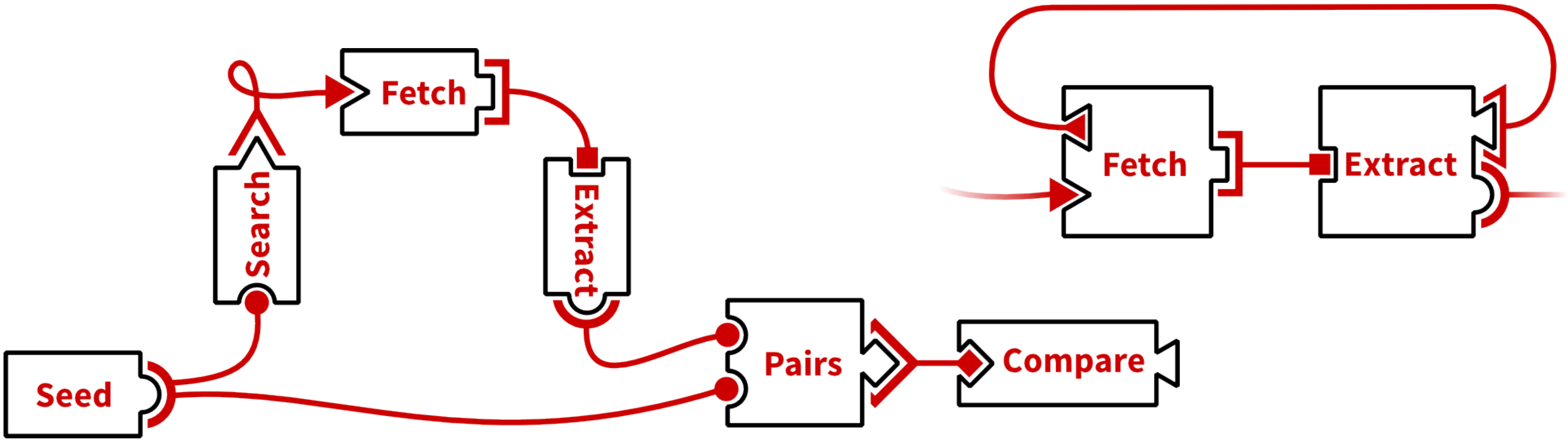
PROJECT OBJECTIVE

So, how to robustly automate this process of gathering, extraction and consolidation of product data?

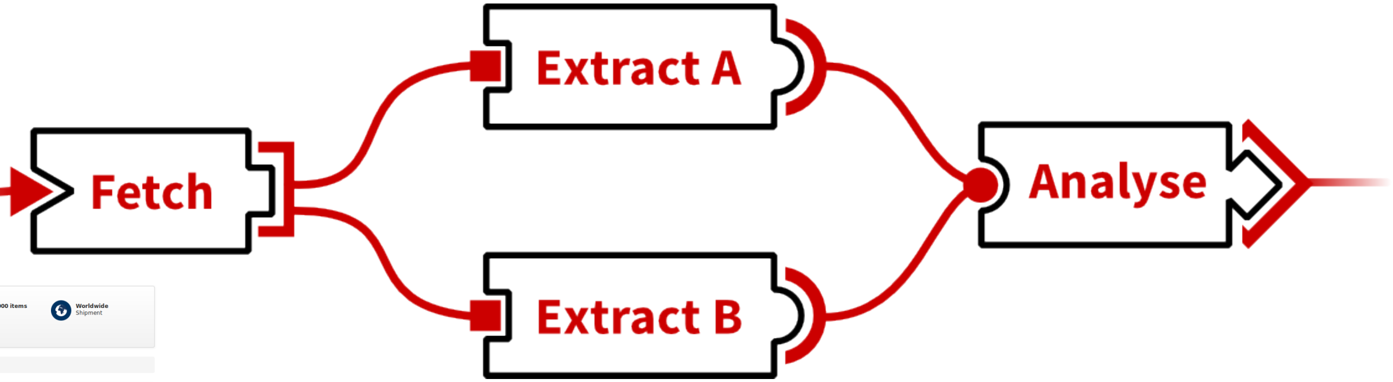
- **Probabilistic approach** throughout
- **Architecture for web harvesting**
  - Automatically understand search forms and page structures, extract fields, and handle absurd data and field names
  - Get or automatically produce feedback to decide about whether something is good or rubbish
  - Be capable of backing out of a decision to redo something

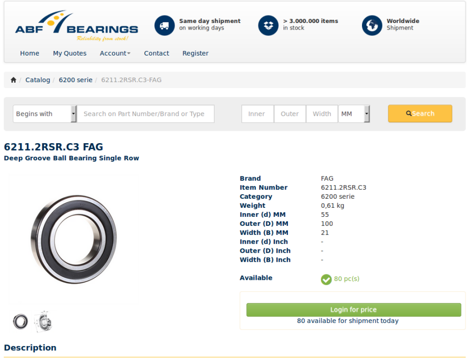
WEB HARVESTING ARCHITECTURE

- Flexible and intelligent
- Backpedal and Redo (data provenance)
- Flows may try multiple methods, sort out results later
- Feedback loops to learn from ‘probably good’ data to understand new sites



PROBABILISTIC THROUGHOUT





ABF BEARINGS *Reliability from stock!*

Same day shipment on working days

> 3.000.000 items in stock

Worldwide Shipment

Home My Quotes Account Contact Register

Catalog / 6200 serie / 6211.2RSR.C3-FAG

Begins with | Search on Part Number/Brand or Type | Inner | Outer | Width | MM | Search

6211.2RSR.C3 FAG

Deep Groove Ball Bearing Single Row

Brand	FAG
Item Number	6211.2RSR.C3
Category	6200 serie
Weight	0,61 kg
Inner (d) MM	55
Outer (D) MM	100
Width (B) MM	21
Inner (d) Inch	-
Outer (D) Inch	-
Width (B) Inch	-
Available	✓ 80 pc(s)

Login for price

80 available for shipment today

Description

JudgeD

Probabilistic DataLog

Extract A

`Width` = `10` [x=1]

`Inner` = `17 mm` [Y=1]

Extract B

`Width` = `12` [x=2]

`Inner` = `See Diagram` [Y=2]

CONCLUSIONS

Goal: Enrich and clean product data

Approach

- Gather and extract from websites
- Consolidate data of individual products

Solution

- Intelligent and flexible architecture for web harvesting
- Probabilistic approach throughout

Repository

- https://github.com/utdb/combine
  Note: academic code — might explode during use

**If a man will begin with certainties,
he shall end in doubts;
but if he will be content to begin with
doubts, he shall end in certainties.**

(Francis Bacon, 1605)

**Doubt is one of the
names of intelligence**

(Jorge Luis Borges, 1979)