

# Creation of a Data-Driven Algorithm to Reduce Cataloging Backlogs

Prototype for CRL's Cooperative Cataloging Initiative

**Edidiong David Ibokete**

Algorithm & Systems Engineer (Intern)



# Background - why a Cooperative Exchange?

- Cataloging backlogs across institutions
- Specialized expertise is unevenly distributed
- Need for fair, reciprocal workload sharing



Backlogs vs. uneven expertise

# Objectives

*To design a cooperative cataloging exchange by modeling institutional needs and capabilities, developing a matching algorithm, and implementing a prototype to generate match recommendations*



## Model needs & capabilities

Materials, languages, levels, trust, performance hooks



## Scoring logic

Here we consider priorities, skill surplus or deficit, trust metrics, and performance history



## Reciprocity enforcement

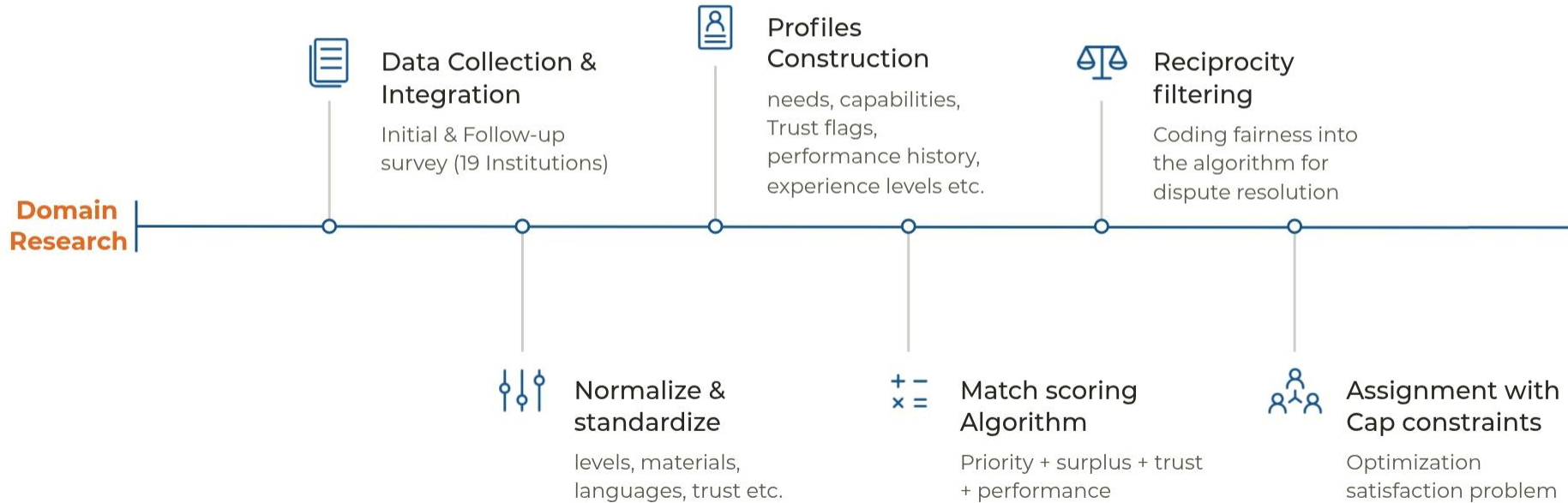
Now we set strict Give and receive caps; capacity caps as well as fairness



## Prototype & test

Python libraries and real survey data from 19 libraries

# Methodology Overview



# Data Collection & Standardization

we gathered data through surveys sent to participating libraries



# Data Standardization & Modeling

- Unified spreadsheet schema

Mapped raw survey fields into consistent, machine-readable columns for the pipeline.

- Needs & Capabilities

Summarize the project and include challenges

- Priorities, trust flags, quantities

Bounded priority scales, PCC/CONSER/NACO indicators, and per-material backlog quantities

- Language Capabilities

Different language capabilities were modeled using Boolean values

- Performance history fields (AUR, AR)

Created placeholder coding for ratings and reliability to weight future rounds (defaults: AUR=0, AR=10)

| CatMat_ LevAcptd:Govt_docs | CatMat_ LevCon:Archival_materials |
|----------------------------|-----------------------------------|
| PCC/Bibco Standard         | RDA Core                          |
| PCC/Bibco Full             |                                   |
| PCC/Bibco Full             | PCC/Bibco Full                    |
| RDA Core                   |                                   |
| RDA Core                   | PCC/Bibco Standard                |
| Minimal                    |                                   |
|                            | PCC/Bibco Full                    |
|                            | PCC/Bibco Full                    |
| RDA Core                   |                                   |
| RDA Core                   | RDA Core                          |

| Qty:Digital_media | CatLang_Con:English | CatLang_Con:Arabic |
|-------------------|---------------------|--------------------|
| 12 Yes            |                     | Yes                |
| Yes               |                     |                    |
| 13                |                     | Yes                |
| 12 Yes            |                     |                    |
| Yes               |                     |                    |
|                   |                     | Yes                |
| 35                |                     |                    |
| 12 Yes            |                     |                    |
| 22 Yes            |                     | Yes                |
| 10 Yes            |                     |                    |
| Yes               |                     |                    |

# Profiles Construction & Scoring

*For each institution, we constructed a comprehensive profile that summarizes its needs and capabilities in a structured way*



```

{
  "Institution": "South Dakota State University",
  "Institution_ID": 7,
  "Staff_Size": "1-2",
  "CanContribute": true,
  "Needs": {
    ("digital_media", "english"): {
      "accept_level": "PCC/BIBCO Standard",
      "accept_level_score": 3,
      "material_priority": 1,
      "language_priority": 7,
      "quantity": 35
    }
  },
  "Capabilities": {
    ("govt_docs", "english"): {
      "contribution_level": "PCC/BIBCO Standard", "contribution_level_score": 3
    },
    ("monographs", "english"): {
      "contribution_level": "PCC/BIBCO Standard", "contribution_level_score": 3
    }
  },
  "TrustFlags": { "PCC_BIBCO": false, "CONSER": false, "NACO": false },
  "PerformanceHistory": {
    "AUR_score": null, "AR_score": null,
    "AUR_ratings": [], "Match_Rejections": 0,
    "Successful_Contributions": 0, "Deadlines_Missed": 0
  }
}

```

# Sample Profile Created

- Institution Metadata**  
 Basic info about the institution
- Needs**  
 A dictionary of all requested cataloging tasks.
- Capabilities**  
 A dictionary of all contributions the institution can offer.
- Trust flags**  
 Indicators of expertise level for the institutions.
- Performance History**  
 A placeholder for performance metrics such as Average user Rating (AuR) and Algorithmic Reliability (AR)

# Scoring Logic - Total Score

- **Material priority score**

Highest weights were given to a material type ranked as high priority, and lowest for lower priorities.

- **Language priority score**

Lower priority ranked languages were given fewer scoring points.

- **Skill surplus**

The contributor's cataloging level was computed in comparison to the requesters required level.

- **Trust bonus**

A bonus was added for any institution with any of the trust flags (PCC/BIBCO, CONSER, NACO)

- **Performance bonus (AUR, AR)**

Created placeholder coding for ratings and reliability to weight future rounds (defaults: AUR=0, AR=10)

```
def score_request_against_contributors(requesting_institution, needs_dict, contributor_profiles):  
  
    matches = []  
  
    for (material, language), need_info in needs_dict.items():  
        required_score = need_info.get("accept_level_score")  
        if required_score is None:  
            continue  
  
        # Extract ranked priorities (may be None if not provided)  
        material_priority = need_info.get("material_priority")  
        language_priority = need_info.get("language_priority")  
  
        mat_score = 6 - material_priority if material_priority else 0 # Max = 5  
        lang_score = 8 - language_priority if language_priority else 0 # Max = 7  
        base_score = mat_score + lang_score  
  
        for contributor, profile in contributor_profiles.items():  
            # Prevent self-matching  
            if contributor == requesting_institution:  
                continue # no self-matching  
  
            # Check if contributor supports this (material, language) combo  
            cap = profile.get("capabilities", {}).get((material, language))  
            if not cap:  
                continue  
  
            contributor_score = cap.get("contribution_level_score")  
            if contributor_score is None:  
                continue  
  
            surplus = contributor_score - required_score  
  
            trust_flags = profile.get("trust_flags", {})  
            trust_bonus = sum([  
                3 if trust_flags.get("PCC_BIBCO") else 0,  
                2 if trust_flags.get("CONSER") else 0,  
                2 if trust_flags.get("NACO") else 0  
            ])   
  
            # Add performance bonus (average of AUR and AR)  
            perf = profile.get("PerformanceHistory", {})  
            aur = perf.get("AUR_score") or 0  
            ar = perf.get("AR_score") or 0  
            performance_bonus = (aur + ar) / 2  
  
            # Compute total score  
            total_score = base_score + max(surplus, 0) + trust_bonus + performance_bonus  
            match = {}
```

# Performance History (AUR + AR)

- **Average User Rating (0-5) from requesters**

Highest weights were given to a material type ranked as high priority, and lowest for lower priorities.

- **Algorithmic Reliability (0-10)**

This included completions, deadlines, rejections

- **Round 1 neutral defaults**

AUR=0.0, AR=10.0 (neutral start)

- **perf term in score:  $(AUR + AR) / 2$**

A simple average average was used. Keeping both signals on comparable footing.



```
# Blank performance metrics (can be updated over time)
"PerformanceHistory": {
  "AUR_ratings": [],
  "AUR_score": None,
  "AR_score": None,
  "Match_Rejections": 0,
  "Successful_Contributions": 0,
  "Deadlines_Missed": 0

# performance bonus (average of AUR and AR)
perf = profile.get("PerformanceHistory", {})
aur = perf.get("AUR_score") or 0
ar = perf.get("AR_score") or 0
performance_bonus = (aur + ar) / 2
```

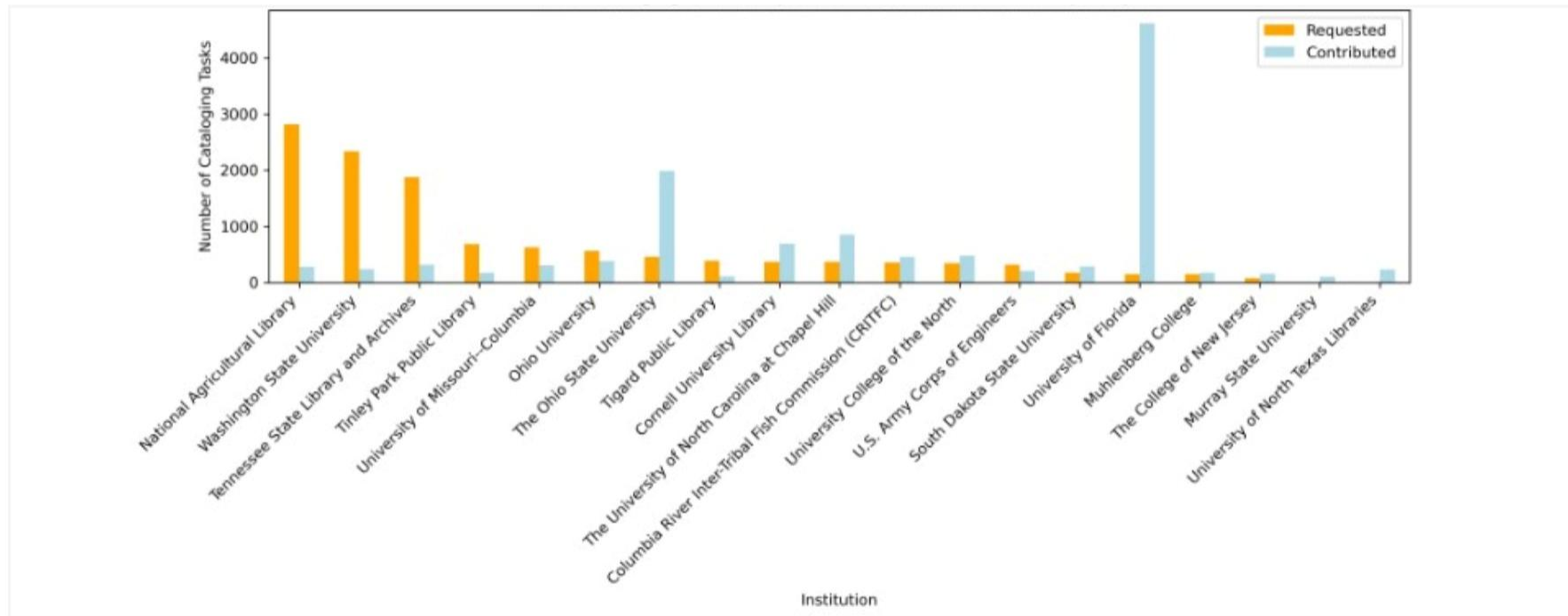
1:1 matches

# Reciprocity and Fairness

*Reciprocity is a first-class rule in this system: Crucially, the exchange was balanced so that the institutions contributed as many tasks as they received*

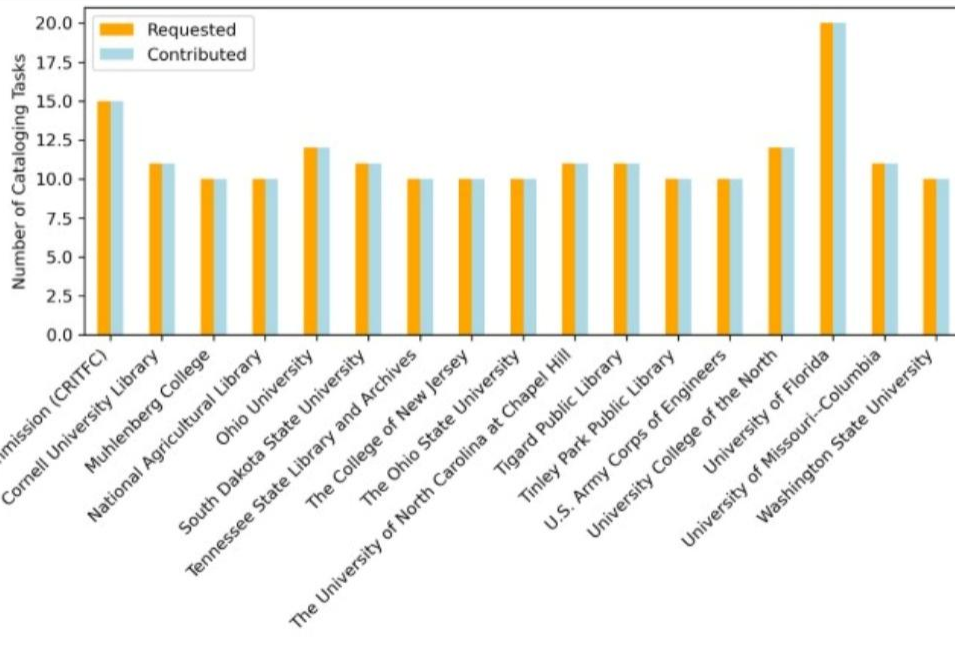


# Raw cataloging tasks: Requested vs Contributed (Pre-reciprocity)



*The chart shows the imbalance in the initial match pool, where some institutions had very high needs but limited contribution capacity, while some others could contribute far more than they requested*

# Cataloging tasks: Requested vs Contributed (Reciprocity enforced)



In the pilot phase, 17 out of 19 organizations were deemed reciprocal, maintaining a balanced distribution of 194 requested and contributed tasks.

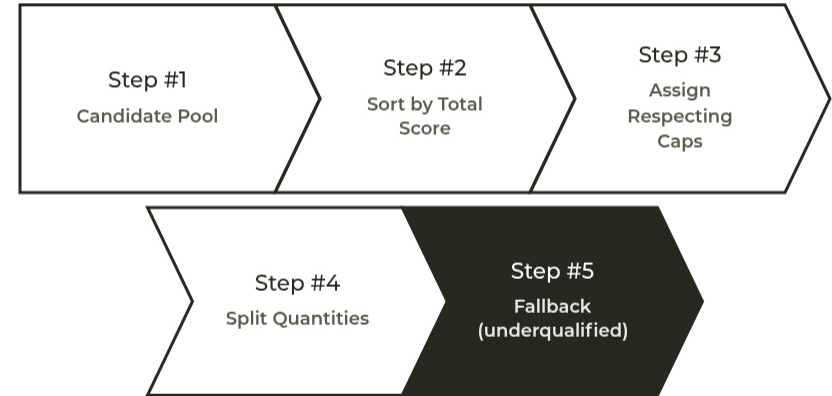
# Assignment & Results

*After running the matching algorithm on the pilot participants of 19 institutions, we achieved feasible exchange assignment.*



# Assignment Pipeline

| requesting_institution                               | contributing_institution                             | material            | language            | assigned_quant | assignment_status       |
|------------------------------------------------------|------------------------------------------------------|---------------------|---------------------|----------------|-------------------------|
| The College of New Jersey                            | The Ohio State University                            | special_collections | arabic              | 10             | Assigned                |
| South Dakota State University                        | Washington State University                          | digital_media       | english             | 10             | Assigned                |
| South Dakota State University                        | Tennessee State Library and Archives                 | digital_media       | english             | 1              | Assigned_Fallback       |
| U.S. Army Corps of Engineers                         | Columbia River Inter-Tribal Fish Commission (CRITFC) | digital_media       | english             | 10             | Assigned_Fallback       |
| Washington State University                          | University of Florida                                | special_collections | arabic              | 10             | Assigned                |
| University of Florida                                | National Agricultural Library                        | govt_docs           | hindi               | 9              | Assigned_Underqualified |
| University of Florida                                | National Agricultural Library                        | govt_docs           | sanskrit            | 1              | Assigned_Underqualified |
| University of Florida                                | University College of the North                      | monographs          | dutch               | 10             | Assigned_Underqualified |
| Cornell University Library                           | Columbia River Inter-Tribal Fish Commission (CRITFC) | monographs          | english             | 5              | Assigned_Fallback       |
| Cornell University Library                           | University of Missouri—Columbia                      | monographs          | english             | 6              | Assigned_Fallback       |
| Ohio University                                      | U.S. Army Corps of Engineers                         | digital_media       | english             | 10             | Assigned_Underqualified |
| Ohio University                                      | University of Florida                                | digital_media       | chinese(simplified) | 2              | Assigned_Fallback       |
| Tinley Park Public Library                           | University of Florida                                | special_collections | arabic              | 8              | Assigned_Underqualified |
| Tinley Park Public Library                           | The University of North Carolina at Chapel Hill      | special_collections | arabic              | 2              | Assigned_Underqualified |
| Muhlenberg College                                   | Ohio University                                      | special_collections | english             | 10             | Assigned                |
| National Agricultural Library                        | Ohio University                                      | digital_media       | english             | 2              | Assigned                |
| National Agricultural Library                        | The College of New Jersey                            | digital_media       | japanese            | 8              | Assigned_Fallback       |
| University College of the North                      | Tennessee State Library and Archives                 | digital_media       | english             | 9              | Assigned_Fallback       |
| University College of the North                      | University of Missouri—Columbia                      | monographs          | english             | 3              | Assigned                |
| Tigard Public Library                                | University of Missouri—Columbia                      | monographs          | english             | 2              | Assigned                |
| Tigard Public Library                                | South Dakota State University                        | monographs          | english             | 9              | Assigned                |
| Columbia River Inter-Tribal Fish Commission (CRITFC) | South Dakota State University                        | govt_docs           | english             | 2              | Assigned                |
| Columbia River Inter-Tribal Fish Commission (CRITFC) | Muhlenberg College                                   | monographs          | english             | 10             | Assigned                |
| Columbia River Inter-Tribal Fish Commission (CRITFC) | Tinley Park Public Library                           | monographs          | english             | 3              | Assigned_Fallback       |
| University of Missouri—Columbia                      | Cornell University Library                           | monographs          | chinese(simplified) | 11             | Assigned_Underqualified |
| Tennessee State Library and Archives                 | The College of New Jersey                            | digital_media       | japanese            | 2              | Assigned                |
| Tennessee State Library and Archives                 | Tinley Park Public Library                           | monographs          | english             | 7              | Assigned                |
| Tennessee State Library and Archives                 | University College of the North                      | monographs          | dutch               | 1              | Assigned_Underqualified |
| The Ohio State University                            | University College of the North                      | monographs          | hindi               | 1              | Assigned_Fallback       |
| The Ohio State University                            | The University of North Carolina at Chapel Hill      | special_collections | farsi_persian       | 9              | Assigned_Fallback       |
| The University of North Carolina at Chapel Hill      | Tigard Public Library                                | monographs          | japanese            | 11             | Assigned                |



# Pilot Results (Numbers)

17

## Reciprocal institutions out of 19

2 institutions were excluded because they had no needs and only offered help (failed reciprocity test)

31

## Assignments completed

A total of 31 match assignments were generated in one experimental run

194

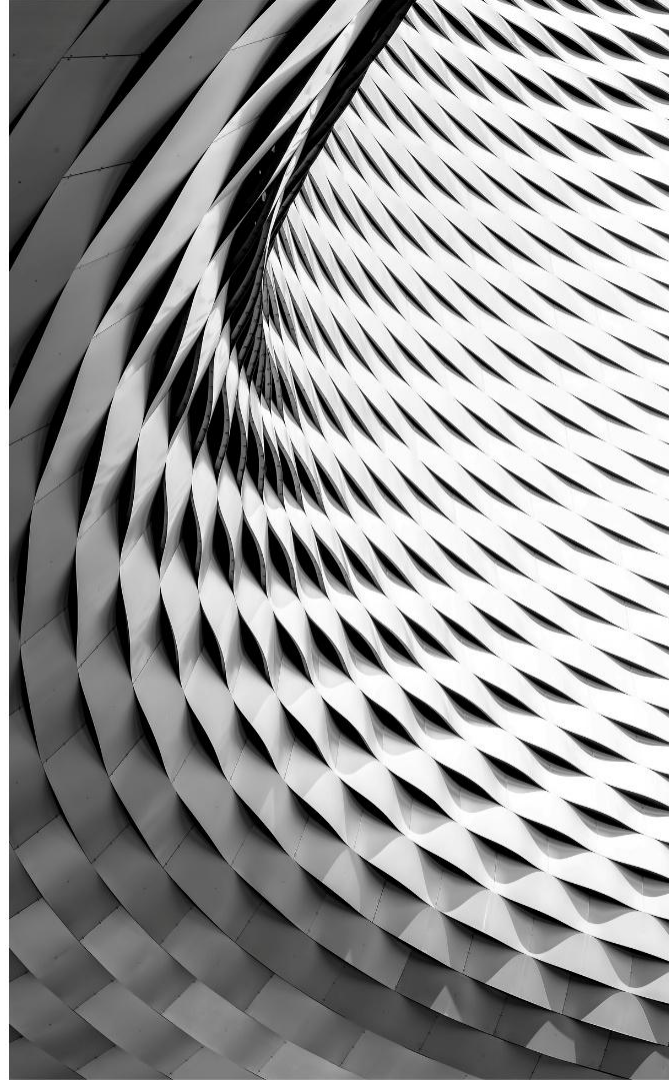
## Tasks matched

Across all 31 matches, 194 cataloging tasks were allocated to be exchanged

100 %

## Reciprocity compliance

194 tasks exactly equals the total number of tasks that the 17 institutions had requested.



# Challenges & Mitigations

During the lifespan of the project, several challenges were encountered

| Challenges |                                         | Mitigations                      |
|------------|-----------------------------------------|----------------------------------|
| 1          | Inconsistent data                       | Normalization dictionary mapping |
| 2          | Assignment optimization                 | Greedy score optimization        |
| 3          | Incorporating performance metrics       | Placeholder architecture         |
| 4          | Imbalance in requests and contributions | Fairness constraint enforcement  |
| 5          | Complex matching criteria               | Heuristic weight calibration     |