Data Engineering Term Project Proposal

Academic Conference Assistant

Team 3

Junyeop Ha **Jinwoo Lee** **Jaesung Kim** **Jungjae Lee**

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Motivation

- **Why Are Conferences Important for Graduate Students?**
  - **Research Development**
    - Present research and receive feedback
    - Discover new research topics and ideas
  - **Researcher Networking**
    - Connect with researchers sharing similar interests
    - Explore opportunities for future collaboration
  - **Career Development**
    - Explore academic and industry career opportunities









Motivation

- **Finding relevant sessions**
  - Too many sessions to review within limited time
- **Identifying relevant researchers**
  - Difficult to find researchers with similar interests
- **Initiating meaningful conversations**
  - Uncertainty about whom to approach and what to discuss
- **Maintaining professional connections**
  - Difficult to track contacts and follow up after conferences

A conference-specific networking assistant for researchers is needed



Parallel Session (Nov. 1st , 8:00-9:15)
SA09 - Healthcare Delivery Operations and Clinical Analytics
SA15 - AI for Forecasting, Manufacturing, and Supply Chaings
SA16 - AI-Enabled Process Improvement in Service Operations
SA36 - Industrial AI: Monitoring, Prognostics, and Predictive Maintenance
SA40 - Data-Driven Approaches to Address Social Disparities
SA42 - AI Applications in Forecasting and Consumer Interaction
SA53 - Resilience Planning and Public Sector Operations
SA59 - Power Grid Operations, Storage, and Large Loads
SA60 - Electricity Market Design and Emerging Grid Technologies
SA70 - Analytics and OR in Higher Education: Curriculum and Pedagogy

Challenges Faced by Conference Attendees

Researcher	A	B	C	D	E	F	G	H	I
Similarity	0.87	0.32	0.31	0.21	0.56	0.10	0.62	0.95	0.72



Parallel Session (Nov. 1st , 8:00-9:15)	Similarity
SA09 - Healthcare Delivery Operations and Clinical Analytics	0.37
SA15 - AI for Forecasting, Manufacturing, and Supply Chaings	0.82
SA16 - AI-Enabled Process Improvement in Service Operations	0.14
SA36 - Industrial AI: Monitoring, Prognostics, and Predictive Maintenance	0.65
SA40 - Data-Driven Approaches to Address Social Disparities	0.91
SA42 - AI Applications in Forecasting and Consumer Interaction	0.28
SA53 - Resilience Planning and Public Sector Operations	0.53
SA59 - Power Grid Operations, Storage, and Large Loads	0.76
SA60 - Electricity Market Design and Emerging Grid Technologies	0.09
SA70 - Analytics and OR in Higher Education: Curriculum and Pedagogy	0.44

Key function

- **Researcher Matching**
- **Conversation Assistance**
- **Networking Management**
- **Session Recommendation**

Data Challenges: Limitations of Existing Researcher Data

Identity ambiguity

Program **J. Park**

ORCID **Jihoon Park**



The same researcher may appear under different names or identifiers across sources.

OpenAlex

Author | Claim | Add to collection

Kangbok Lee

Observed names: Kang-Bok Lee, LEE, KANGBOK, Lee, Kangbok

Institutions: Pohang University of Science and Technology, Korea Post, Korea Institute of Industrial Technology

Observed institutions: Rutgers, The State University of New Jersey, University of Science and Technology, Pohang University of Science and Technology, Soongsil...

show all 13

ORCID: 0000-0002-3526-9865

Observed ORCIDs: 0000-0002-3526-9865 0000-0001-9508-5577

OpenAlex ID: A5038840736



Name: Tom Cruise
Born In: 1962
Birthplace: New York



Name: Thomas Cruise
Birthday: July 3, 1962
Birthplace: NY



Name: T. Cruise
Job: Actor
Birthplace: USA



Name: T. C.
Job: Actor, Producer
Birthplace: Syracuse, NY

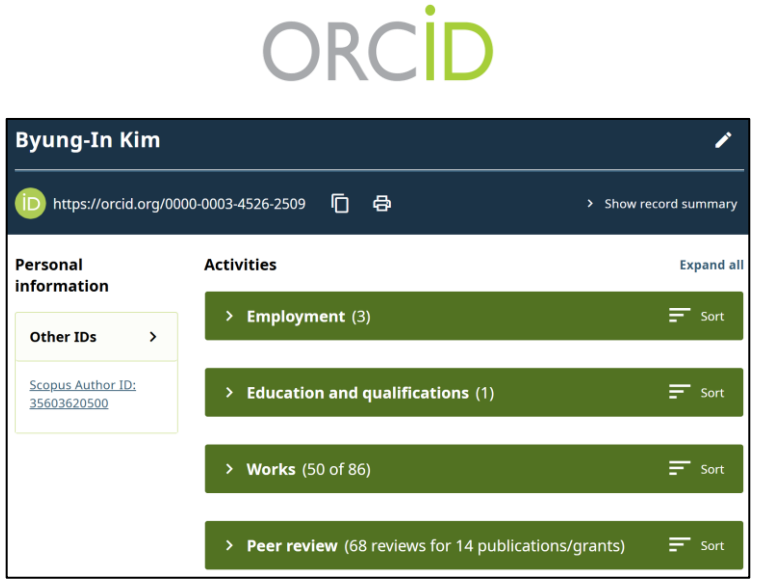
→ **Cross-validate researcher identities using multiple identifiers and profile attributes.**

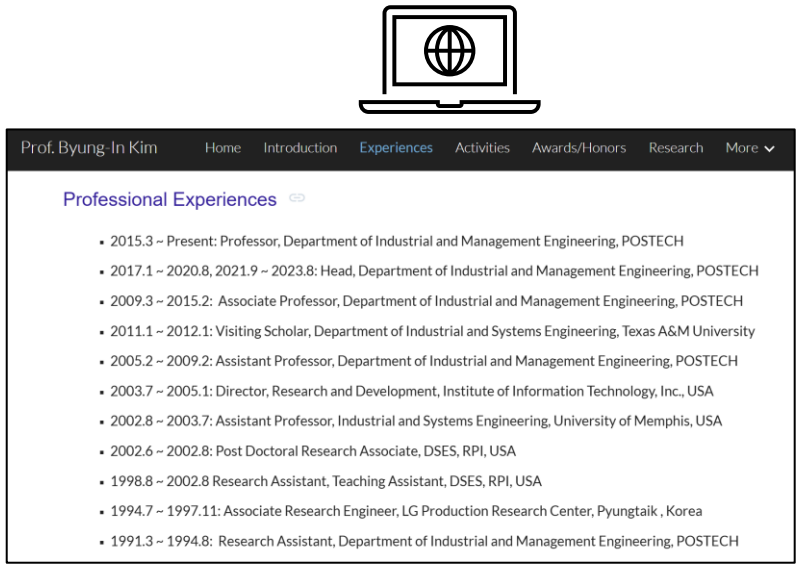
Data Challenges: Limitations of Existing Researcher Data

Incomplete & sparse data

	ORCID	Scholar	Program
Affiliation	POSTECH	KAIST	POSTECH
Papers	3	27	n/a
Abstract	n/a	n/a	missing

Researcher information may be incomplete or sparse across sources.





→ **Integrate data from multiple sources to fill gaps and identify inconsistencies**

Data Challenges: Limitations of Existing Researcher Data

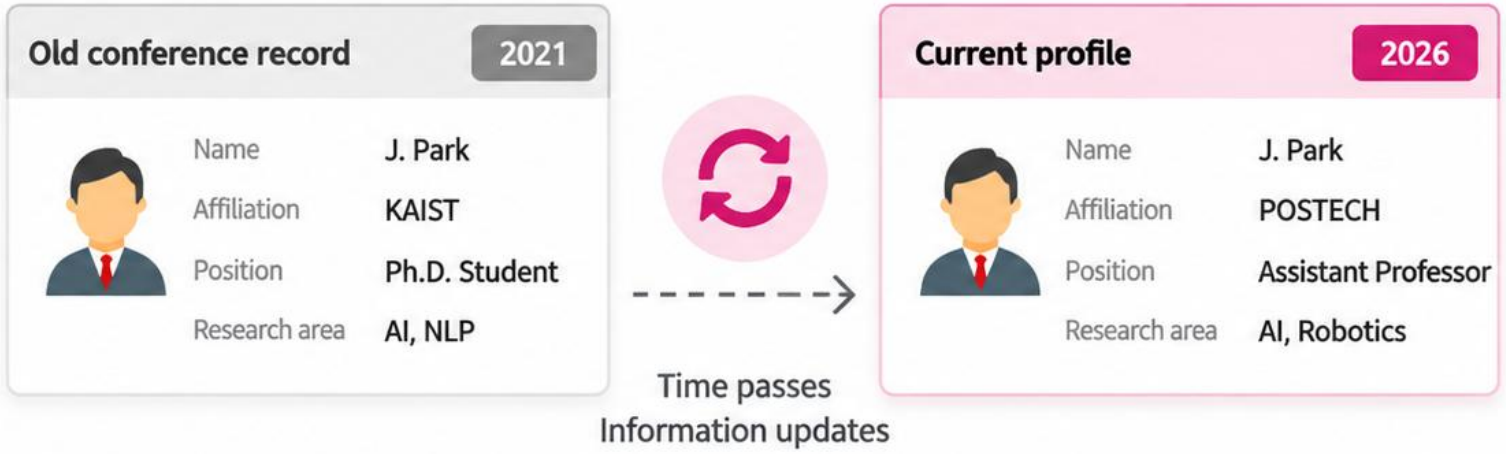
Outdated records

KAIST
2021

→

POSTECH
2026

Researcher affiliations and profiles change over time. Researcher information is not static.



→ **Keep researcher profiles update and automatically detect changes**

Overview

INPUTS

Conference event / program

https://meetings.informs.org/

HTTPX / Requests

Researcher raw data

ORCID API

User info

Web form / app input

Event-triggered

Pipeline A Conference-driven Expansion

New conference → Extract sessions & presenters → Resolve identities & link people → Validate & upsert new / existing

Integrated Academic Database

Persistent records shared across conferences

- Conference + Sessions
- Presentations
- Researchers + Papers
- User + Network

Raw snapshots | Source IDs | Matching status | Sync timestamps

Monthly

Pipeline B Periodic Researcher Enrichment

Select tracked researchers → Re-fetch ORCID records → Detect changes → Incremental update + feature refresh

OUTPUTS

Generate features

Research profile + session text + user preferences

Similarity / ranking → LLM context

Conference-time services

01 Relevant sessions

02 Researchers to meet

03 Conversation starters

04 Connection follow-ups

Service Application

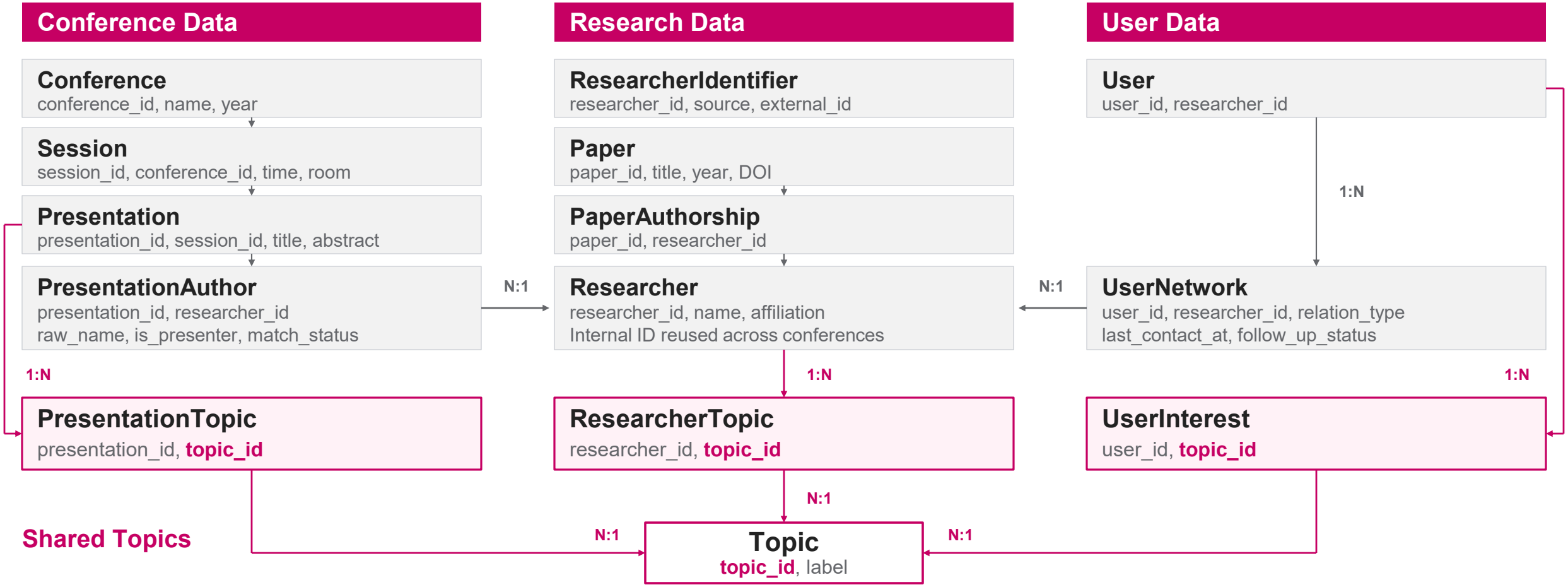
FastAPI

Streamlit

AI Assistance

Database schema

One PostgreSQL database shared across conferences



Provenance & Sync Metadata

RawRecord: source, source_id, payload, fetched_at SyncState: researcher_id, source, last_success_at, status

Data collection pipeline

Pipeline A Conference-driven Expansion

Runs: **a new conference is added**

1 Collect program	2 Resolve identities	3 Enrich researchers	4 Save records
Sessions, talks, and presenters	Link names to known researchers, or add new ones	Papers and profile from ORCID	Remove duplicates, keep unmatched names

Pipeline B Periodic Researcher Enrichment

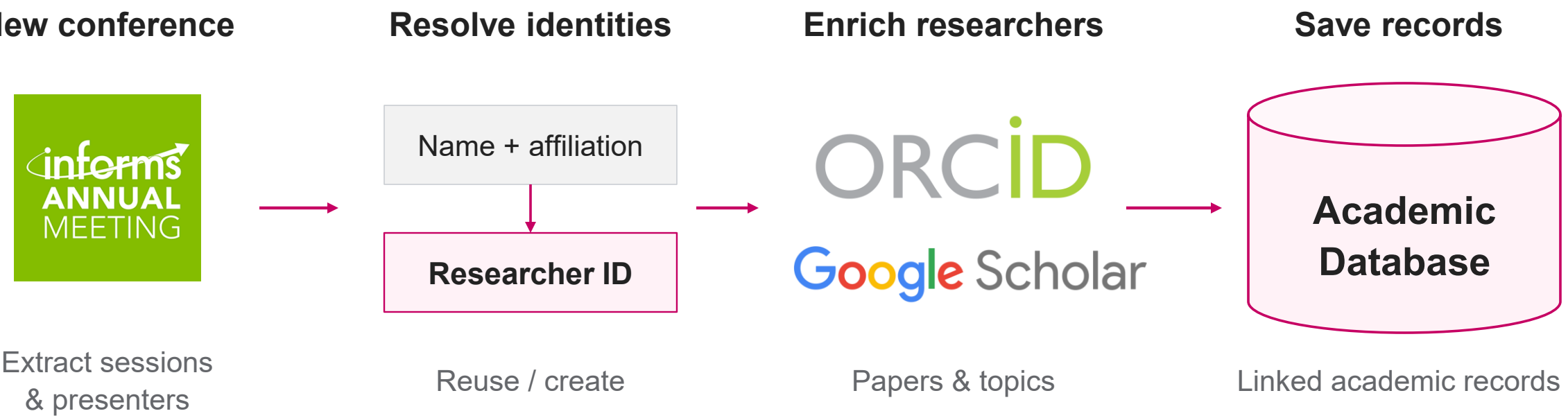
Runs: **every month**

1 Pick Tracked researchers	2 Re-fetch ORCID	3 Detect changes	4 Update & Refresh
From past conferences or followed by users	Fetch the latest public records	Compare with the last saved version	Save changes, refresh topics

User input Users add their profile, research interests, and contacts through the web app.

Pipeline A Conference-driven Expansion

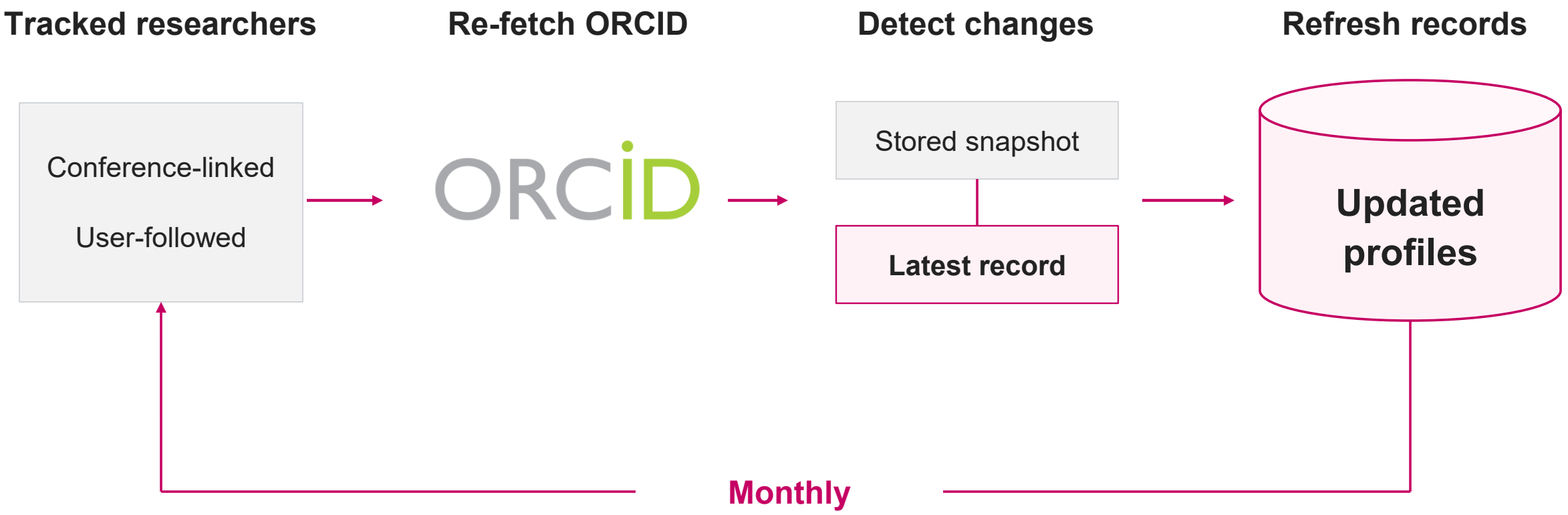
Triggered by a new conference program



Existing researchers keep the same ID across conferences

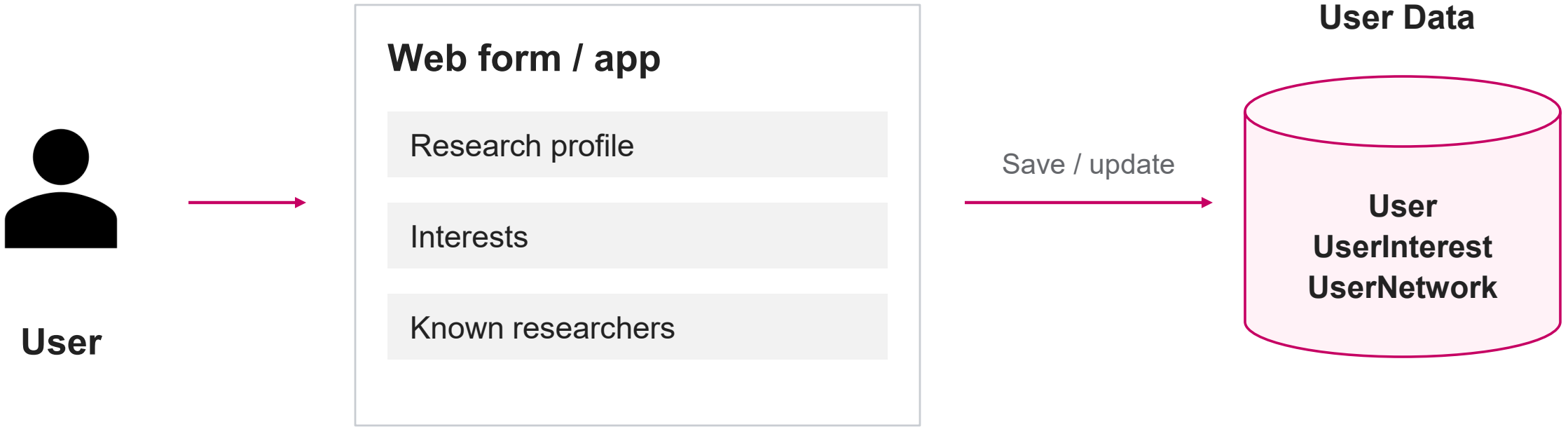
Pipeline B Periodic Researcher Enrichment

Keep existing researcher records current



User Input

Personal information collected through the app



Later updates: contact notes + follow-up status

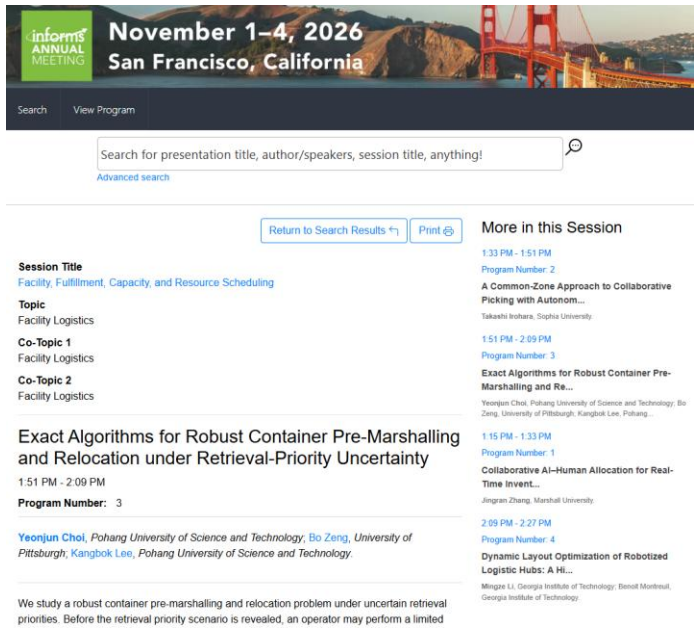
Data Ingestion & Transformation | INFORMS Annual Meeting 2026

From conference webpages to normalized schedules and presentation records

01 RAW DATA

INFORMS 2026 Program

HTML / itinerary



02 EXTRACT & TRANSFORM

01 FETCH & SNAPSHOT

Collect program HTML (render if needed); retain page URL, HTML, and fetched_at.

↓

02 PARSE & NORMALIZE

Extract session code, title, date/time, track and room; normalize timestamps.

↓

03 VALIDATE & UPSERT

Enforce unique conference + session; upsert changes and cancellations.

03 STRUCTURED DATABASE

Conference
conference_id, name, year

Session
session_id, conference_id, time, room

Presentation*
presentation_id, session_id, title, abstract

PresentationAuthor*
presentation_id, researcher_id
raw_name, is_presenter, match_status

PresentationTopic* / Topic
presentation_id, **topic_id**; label

Topic links derived from talk titles / abstracts.
** Talk / author details only when accessible.*

DATA QUALITY & TRACEABILITY Every record keeps the raw page and download time, so schedule changes can be tracked.

Data Ingestion & Transformation | ORCID

From nested researcher records to normalized academic entities

01 RAW DATA

ORCID Public API v3.0

JSON /record


```
{
  "orcid": "0000-0002-3526-9865",
  "name": "Kangbok Lee",
  "affiliation": {
    "institution": "POSTECH",
    "department": "Industrial and Management Engineering",
    "position": "Associate Professor"
  },
  "scopus_author_id": "24437178700",
  "works": [
    {
      "title": "The circular balancing problem",
      "year": 2025,
      "journal": "European Journal of Operational Research",
      "doi": "10.1016/j.ejor.2024.08.020“
                    ...
```


Public fields: ID · name · affiliations · publications

02 EXTRACT & TRANSFORM

01 FETCH & SNAPSHOT

GET /record; retain JSON, request URL, and retrieval timestamp.

↓

02 PARSE & NORMALIZE

Flatten person, employment, and work groups; standardize ORCID / DOI / dates.

↓

03 VALIDATE & UPSERT

Handle nulls and duplicate works; upsert researcher and publication IDs.

03 STRUCTURED DATABASE

Researcher
researcher_id, name, affiliation

ResearcherIdentifier
researcher_id, source, external_id

Paper
paper_id, title, year, DOI

PaperAuthorship
paper_id, researcher_id

ResearcherTopic / Topic
researcher_id, **topic_id**; label

Topic links derived from profile / work text.
Coauthors may require additional sources.

DATA QUALITY & TRACEABILITY Every record keeps the original JSON and source ID, so any value can be checked later.

Project plan / WBS

Proposal this week: Oct 19–25

ID	Work package	W1 10/19–25	W2 10/26–11/1	W3 11/2–8	W4 11/9–15	W5 11/16–22	W6 11/23–29	W7 11/30–12/6	W8 12/7–13	W9 12/14–15
1	Schema & DB setup	■	■							
2	Conference ingestion (A)		■	■						
3	ORCID enrichment & sync (B)			■	■	■				
4	Similarity & ranking				■	■	■			
5	User app & LLM assistance					■	■	■		
6	Integration & evaluation							■	■	
7	Final demo & presentation								■	■

Checkpoints

Nov 8
Similarity Logic

Nov 22
Linked DB ready

Dec 6
Integrated demo

Dec 15
Final presentation

Expected Benefits

For data

Fill missing information



Researcher info that conference sites don't show

More reliable data



Checked records, more accurate than sources like OpenAlex

For users

Personal recommendations



Research areas matched to each user's interests

For the research community

Stronger networking



Better research outcomes and visibility

Global collaboration



Easier communication across countries

“From scattered records to meaningful connections.”

Thank you