The ILL Publication Matcher (PUMA)

FILL2030 WP3

Summary

- Objectives of the PUMA project
- Data workflow
- Next steps

Project objectives

Data retrieval and exploitation

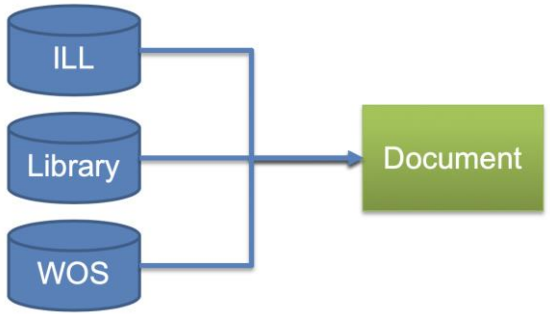
- Development of business intelligence tool for ILL
  - Provide information on who really uses ILL data (not just experimental team)
  - Where to look for new users?
  - Analyse scientific trends
  - Initially designed for the management, aim to support scientists too
- Match publications to experimental data
  - Provide insights on science done and its outcome
  - Challenge to do this in absence of DOIs
- Build corpus of publications and proposals
  - Documents related to ILL or neutron scattering
- Develop tool to analyse publications and match to proposals
  - Data mining and Comparison algorithms

Data workflow: Data Sources

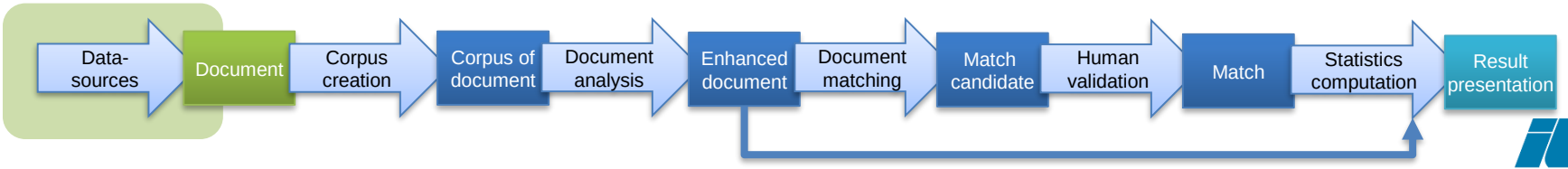
Publication and Proposal retrieval

- Obtain comprehensive corpus of documents related to neutron scattering

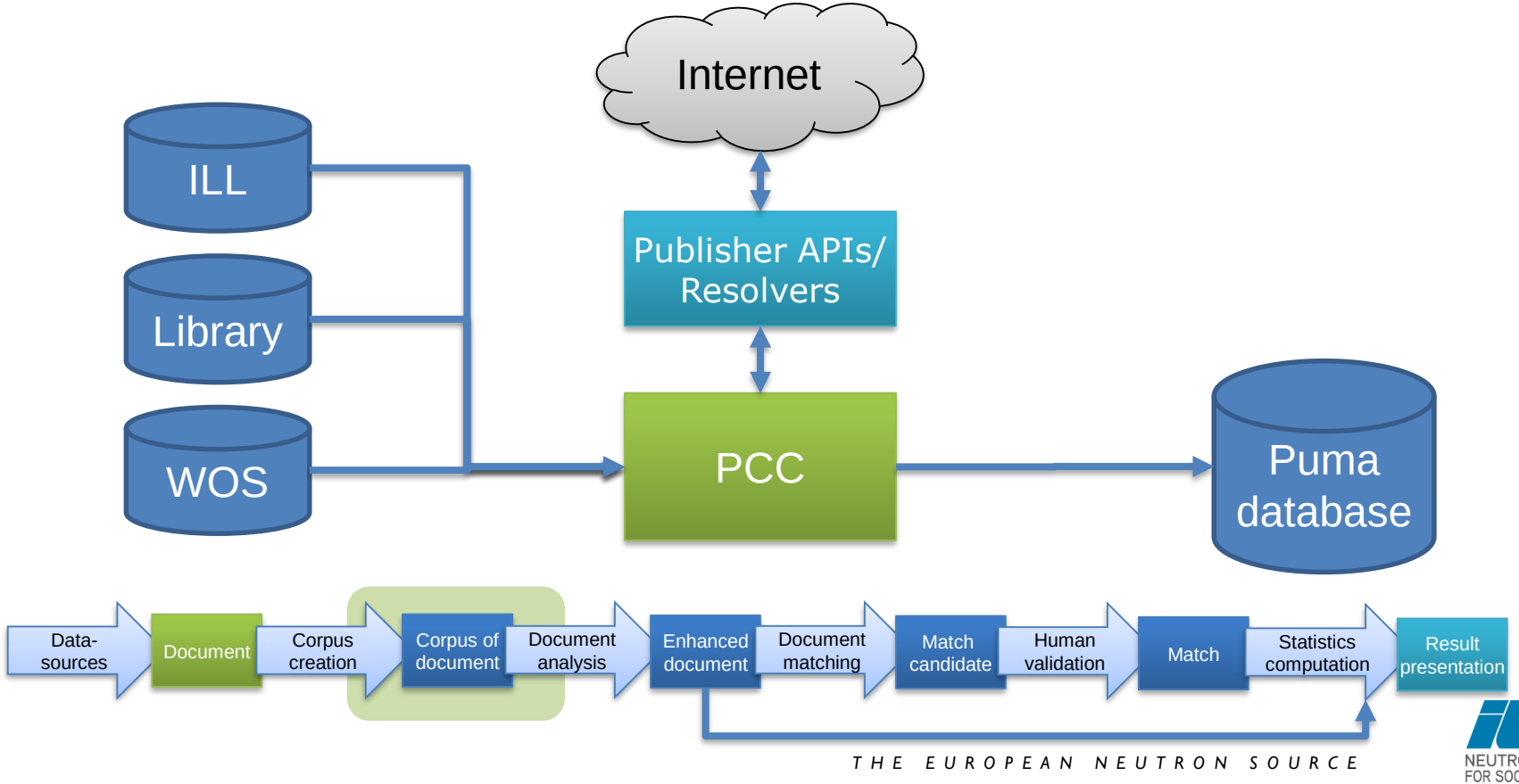
- Three Data sources
  - Web of science
  - ILL Proposal Database
  - Joint ILL/ESRF library



- Full Text not provided by Web of Science
  - Use publisher APIs (Elsevier, APS)
  - Use *resolver* system to determine publisher download URL

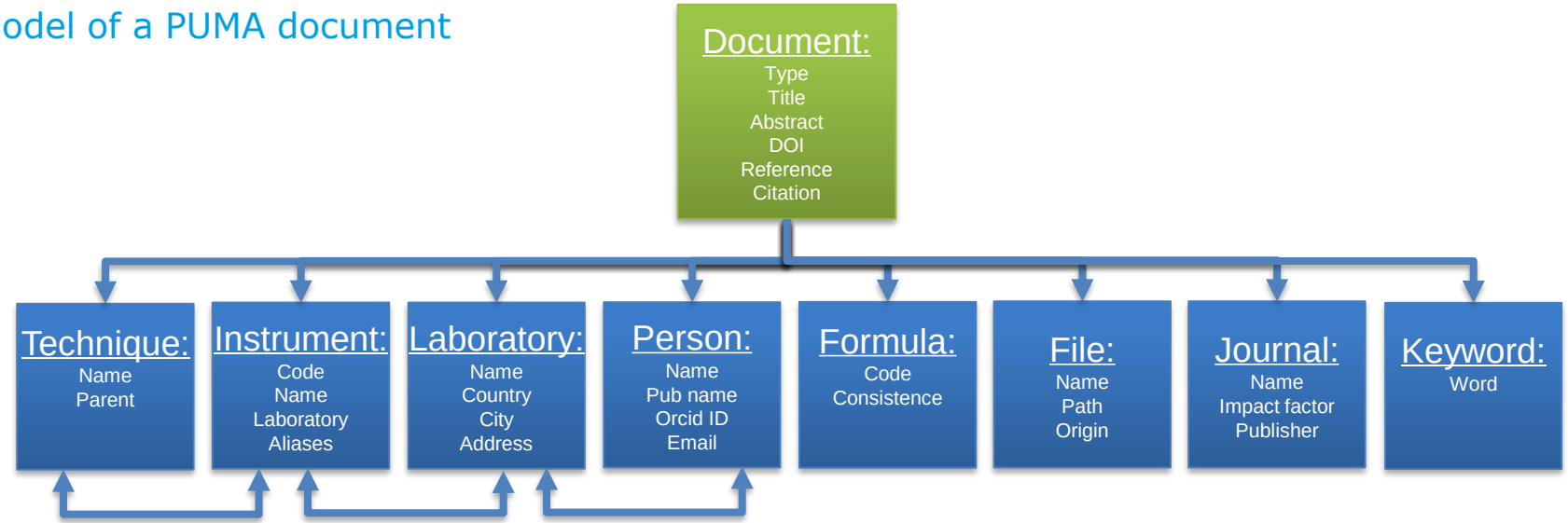


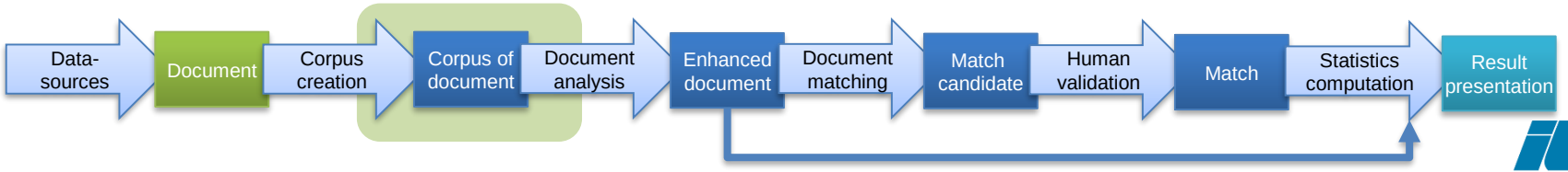
Data workflow: Puma Corpus Creator



Data workflow: Document

Model of a PUMA document

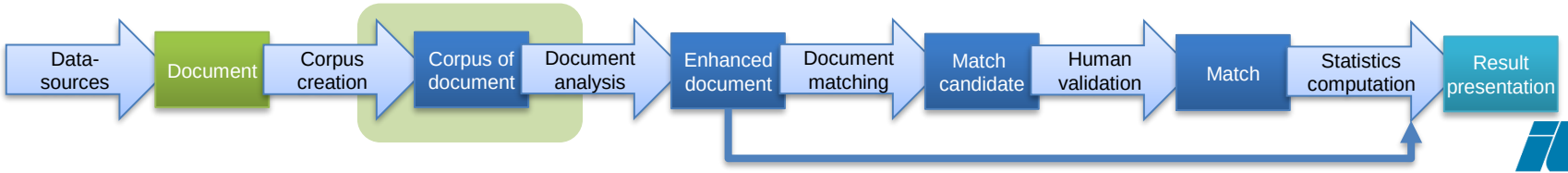




Data workflow: Document

Data provided by the data sources

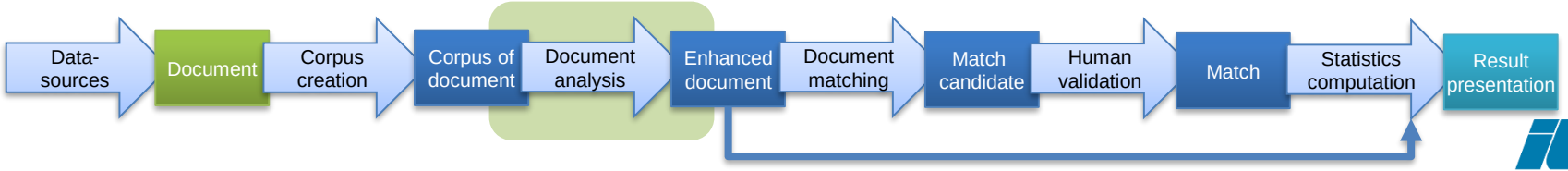
Data-source	Technique	Instrument	Laboratory	Person	Formula	File	Journal	Keyword
WOS			YES	YES		YES*	YES	YES
ILL	YES	YES	YES	YES	YES	YES		
Library		YES	YES	YES		YES*	YES	YES



Data workflow: Data mining

Retrieve missing data from documents

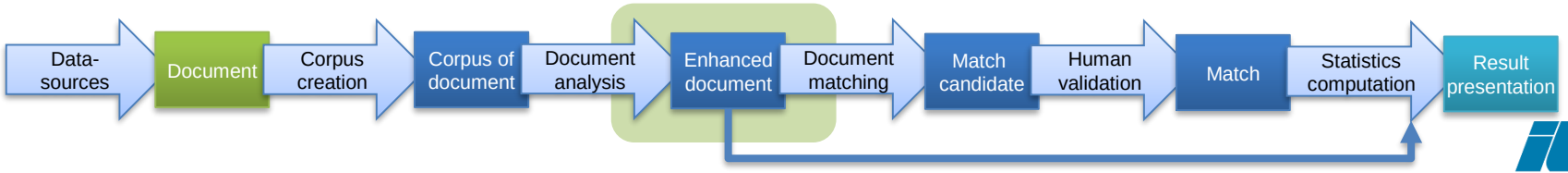
- Extract from full text of :
  - Instruments
  - DOIs
  - Proposals code
  - Chemical formulae
  - Images
- Simplest algorithms use pattern matching
- Most advanced algorithms use artificial intelligence technics



Data workflow: Data mining

Enhanced document metadata

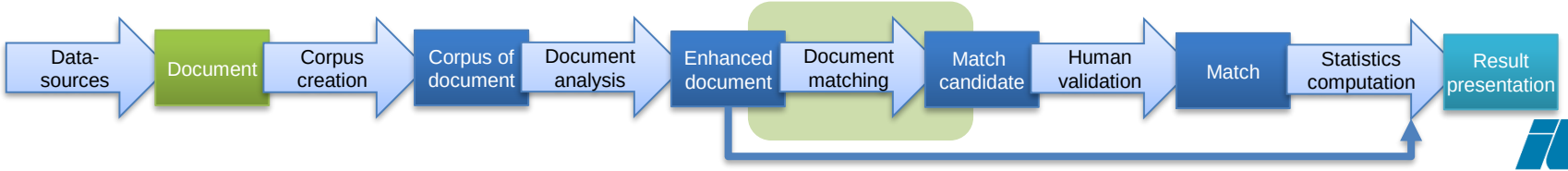
Data-source	Technic	Instrument	Laboratory	Person	Formula	File	Journal	Keyword
WOS	YES	YES	YES	YES	YES	YES*	YES	YES
ILL	YES	YES	YES	YES	YES	YES		
Library	YES	YES	YES	YES	YES	YES*	YES	YES*



Data workflow: Matching

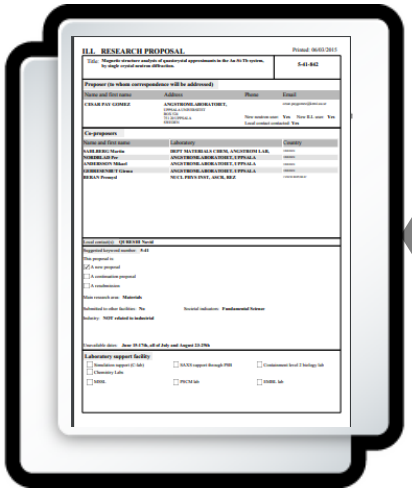
Matcher program: a big data program

- Retrieve links between publications on proposals in absence of DOI
- ~65k Publications & ~19k Proposals
  - >1 billion publication-proposal pairs
  - Document comparisons take several weeks on single core
- Design software and algorithms to handle this amount of data
  - Parallel processing platform: Apache Spark
  - Deduplicate documents and persons to reduce load
  - ~1 hour to perform matching analysis

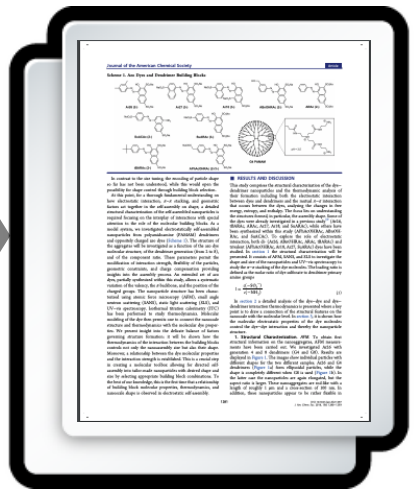


Data workflow: Matching

Data used for matching



Matching of :
DOIs
Proposal codes
Persons
Person laboratories
Keywords
Chemical formulae
Instruments
Document full text
Document abstract
Document title
Images

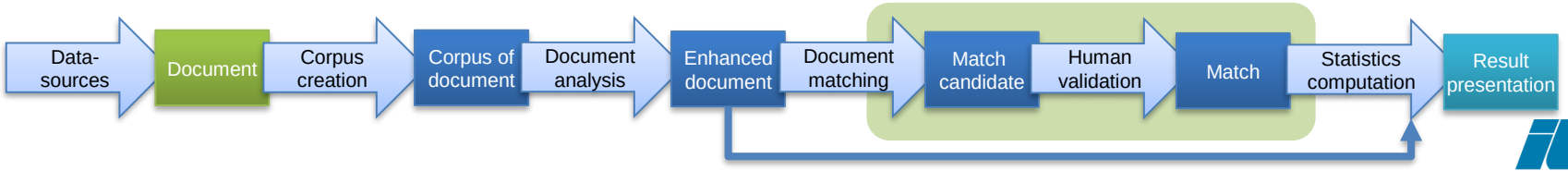


Data-sources → Document → Corpus creation → Corpus of document → Document analysis → Enhanced document → Document matching → Match candidate → Human validation → Match → Statistics computation → Result presentation

Data workflow: Matching

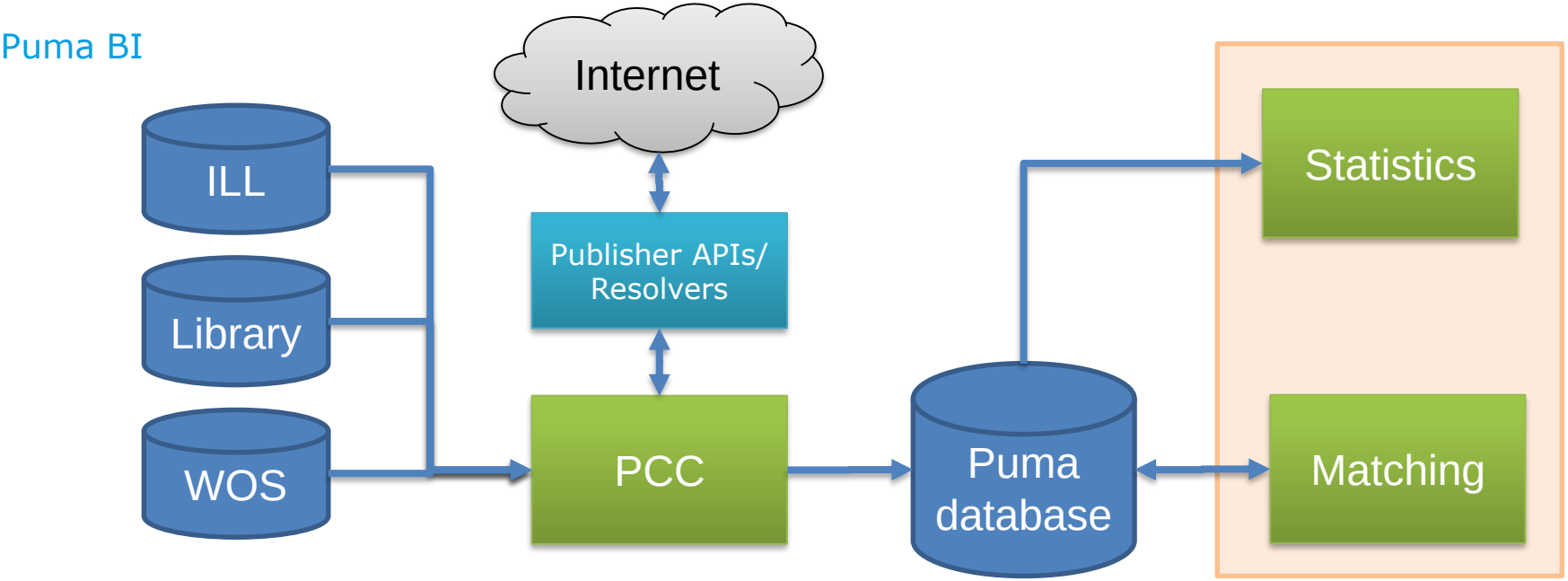
Scoring, optimisation, ranking

- Obtain a number of *candidate matches* for a particular document
  - Each candidate has different *similarities* and *scores*
- Use human validated set of matches to optimise/weight combined scores
  - Produces global score and ranking of each candidate
- >95% confidence that true match is in top 5 candidates
  - Top candidates presented to user for human validation



Data workflow: Data analysis

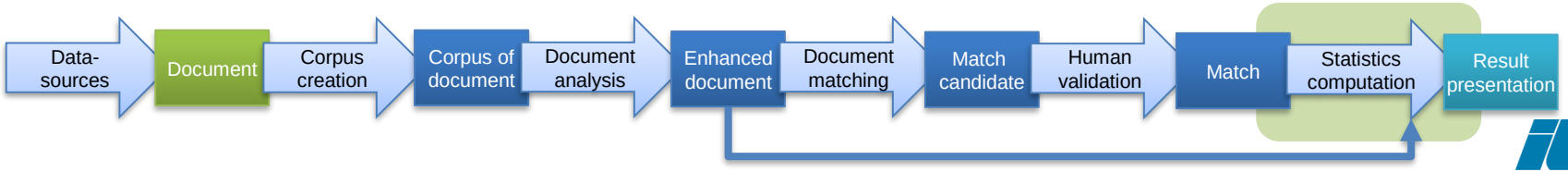
Puma BI



Data analysis

Puma BI statistics from publications and matched documents

- Impact factor of publications using ILL data
  - Determine collaborative impact
- Communities
  - Who used the data
  - Potential new users
- Matches provide new information
  - Analyse delay between experiment and publication
  - Impact of reactor operational days on publications
  - Difference between experimental team and publication authors



Next steps

- Continue developing BI tools: statistics and data visualisation
  - Detect *teams* of authors (impact measure)
  - Detect publications without ILL authors
  - Journal evolutions
- Need to reach full potential of the tool
  - several thousand matches required to obtain statistical significant results
- Collaborative development
  - Open source project (Github repositories)
  - ESRF recruiting to develop Puma
  - In discussion with CERN (similar objectives/requirements)

Further ideas

- Technical improvements to matcher
  - Add matching algorithms
  - Improve trend analysis
- PDF drag and drop and real time analysis
  - Improve workflow for a scientist to produce matches
- Import/analyse supplementary materials
  - DOIs and other information not in principal article
- Try AI algorithms to improve scoring


INSTITUT LAUE LANGEVIN
NEUTRONS
FOR SOCIETY