Building an Indonesian Wordnet

Desmond Dharma Putra
Abdul Arfan
Ruli Manurung
Presented by: Bayu Distiawan

Wordnet

- Useful lexical resource
  - words are clustered into synsets based on their meaning
  - Semantic relationship
- Widely used in IR & NLP Research
- Has been built for many different languages
- In 2008, we started built Indonesian Wordnet

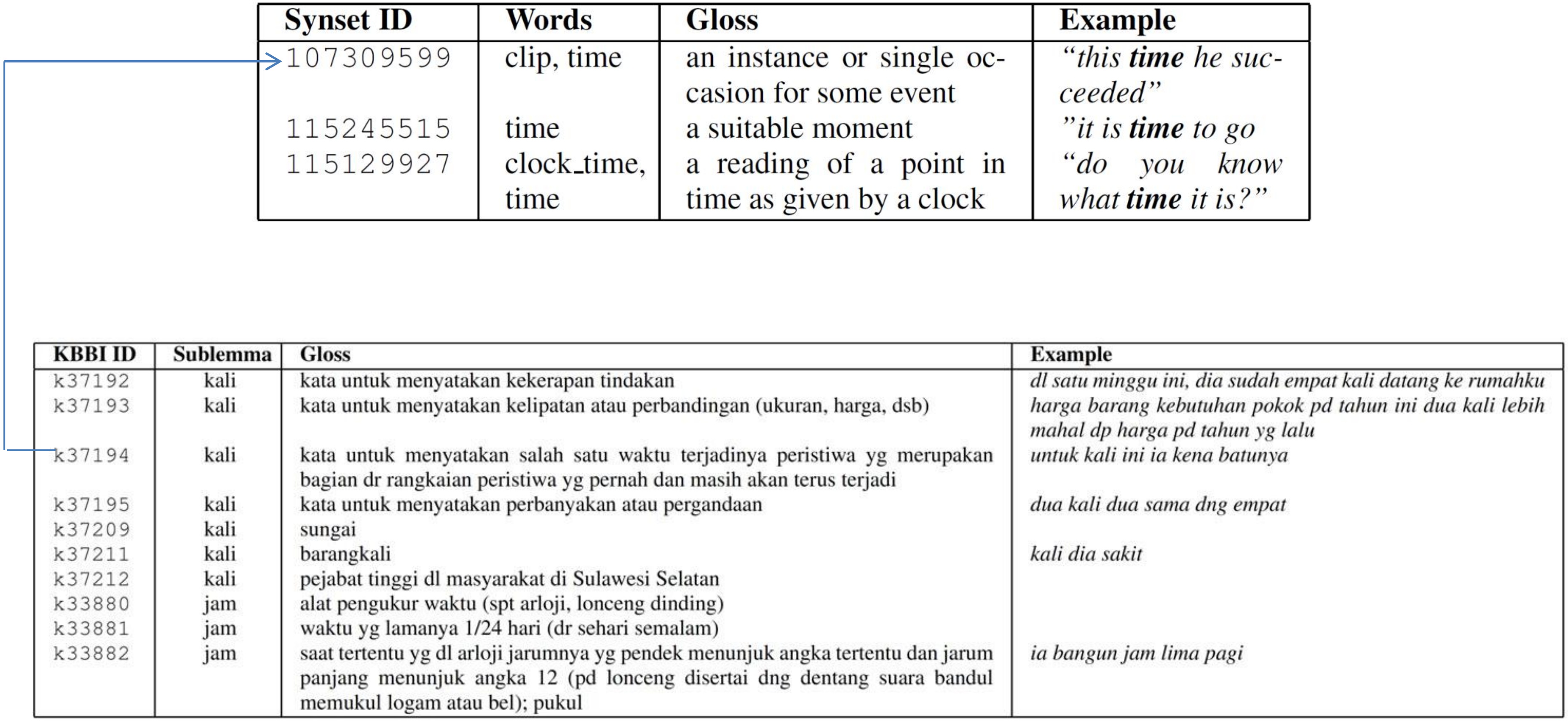
General Aproach on Building Wordnet

- *Expand* model *
  - Synsets in PWN are translated to target language
  - All semantic relation also transferred
  - Easily to built
  - Heavily depend PWN structure
- *Merge* model
  - Specify synset in target model along with their semantic relation
  - Map Synset in target model to synset in PWN
  - Costly to built

Our Approach

- Use *expand* model
  - For every PWN enty --> Find the related word in Bahasa (KBBI)
  - Map word/lemma/sublemma to appropriate PWN synset (concept mapping)

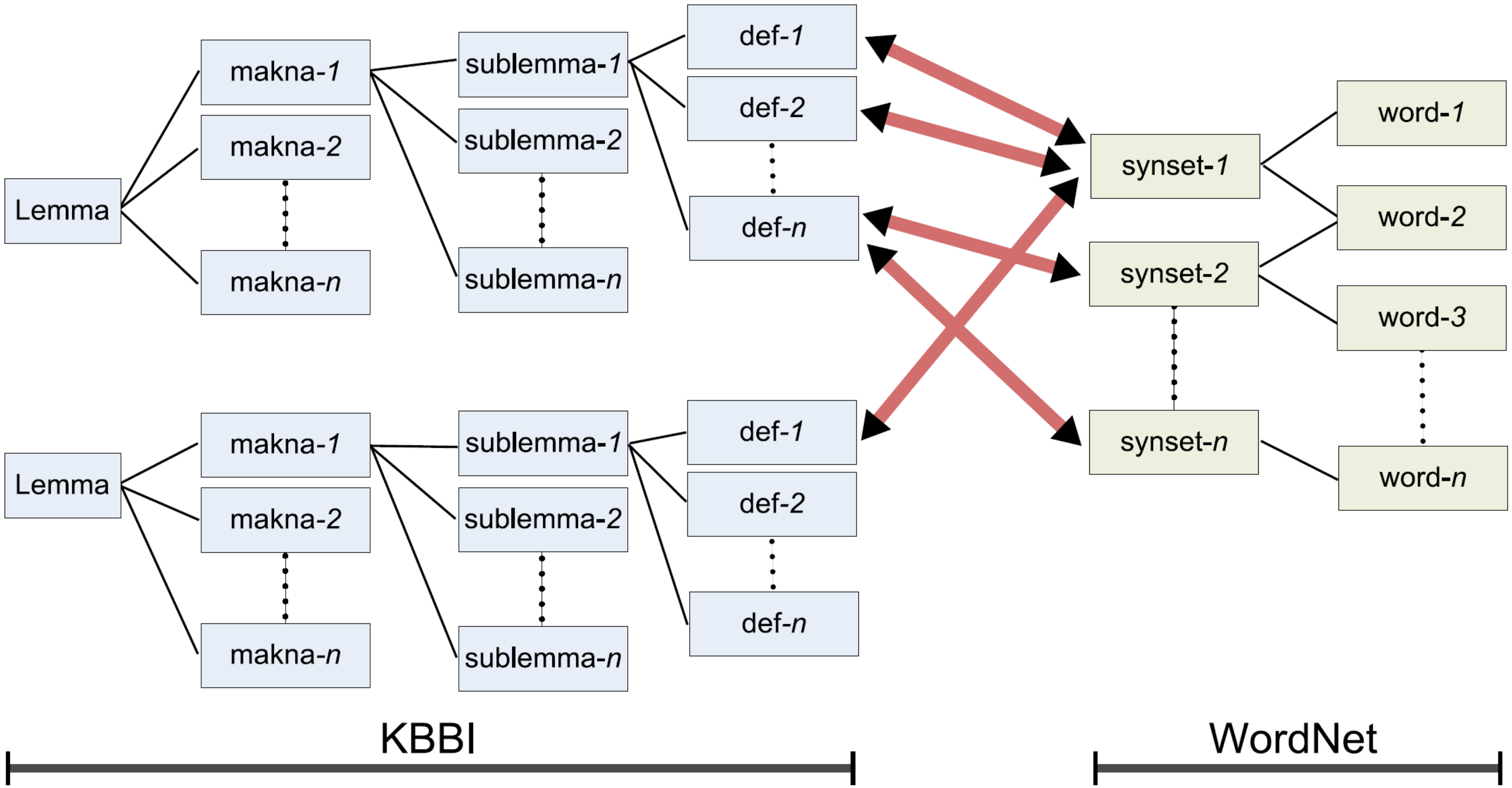
Concept Mapping



Synset ID	Words	Gloss	Example
107309599	clip, time	an instance or single occasion for some event	*“this **time** he succeeded”*
115245515	time	a suitable moment	*”it is **time** to go*
115129927	clock_time, time	a reading of a point in time as given by a clock	*“do you know what **time** it is?”*

KBBI ID	Sublemma	Gloss	Example
k37192	kali	kata untuk menyatakan kekerapan tindakan	*dl satu minggu ini, dia sudah empat kali datang ke rumahku*
k37193	kali	kata untuk menyatakan kelipatan atau perbandingan (ukuran, harga, dsb)	*harga barang kebutuhan pokok pd tahun ini dua kali lebih mahal dp harga pd tahun yg lalu*
k37194	kali	kata untuk menyatakan salah satu waktu terjadinya peristiwa yg merupakan bagian dr rangkaian peristiwa yg pernah dan masih akan terus terjadi	*untuk kali ini ia kena batunya*
k37195	kali	kata untuk menyatakan perbanyakan atau pergandaan	*dua kali dua sama dng empat*
k37209	kali	sungai	
k37211	kali	barangkali	*kali dia sakit*
k37212	kali	pejabat tinggi dl masyarakat di Sulawesi Selatan	
k33880	jam	alat pengukur waktu (spt arloji, lonceng dinding)	
k33881	jam	waktu yg lamanya 1/24 hari (dr sehari semalam)	
k33882	jam	saat tertentu yg dl arloji jarumnya yg pendek menunjuk angka tertentu dan jarum panjang menunjuk angka 12 (pd lonceng disertai dng dentang suara bandul memukul logam atau bel); pukul	*ia bangun jam lima pagi*

Concept Mapping

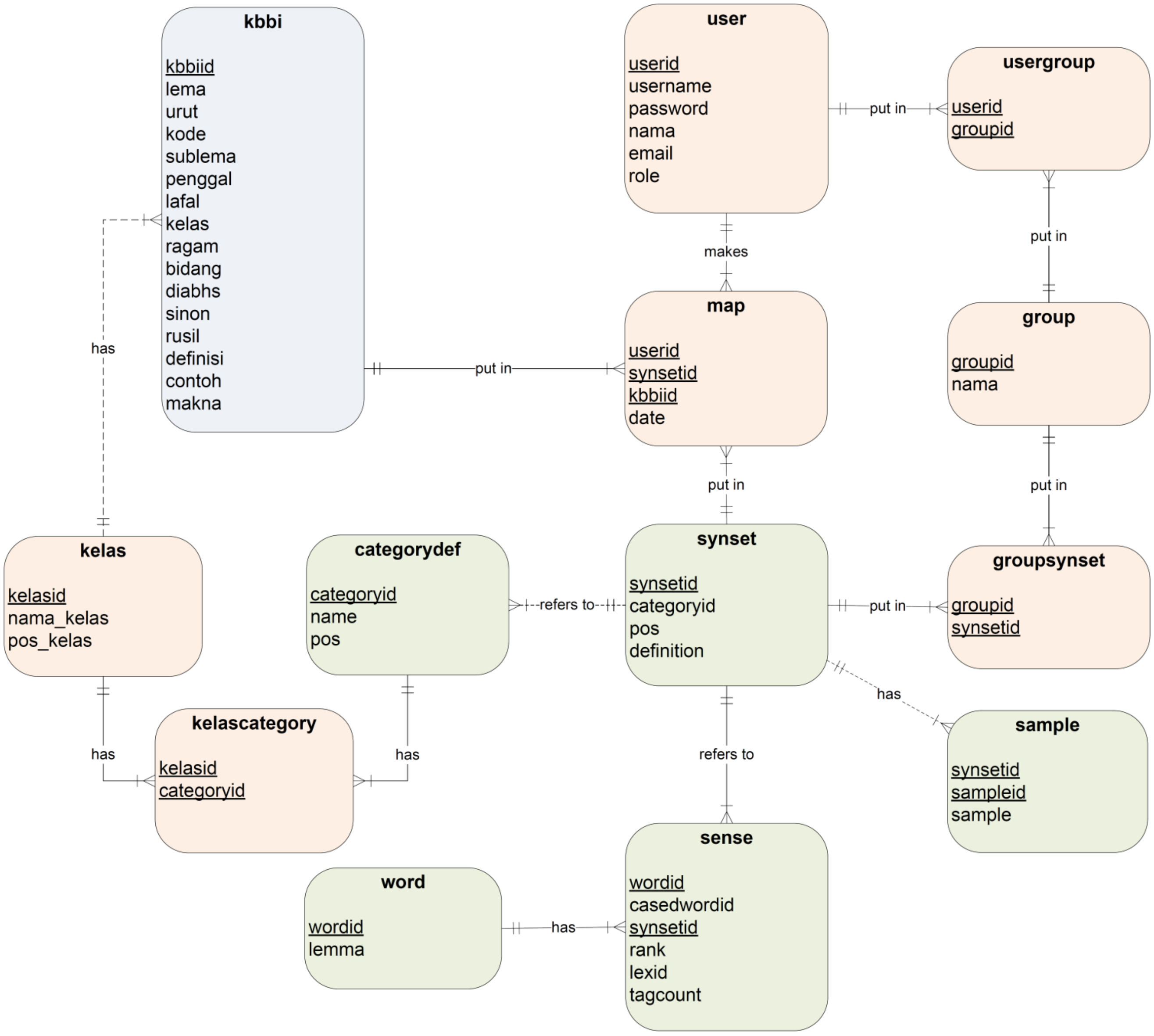


How to do the Mapping?

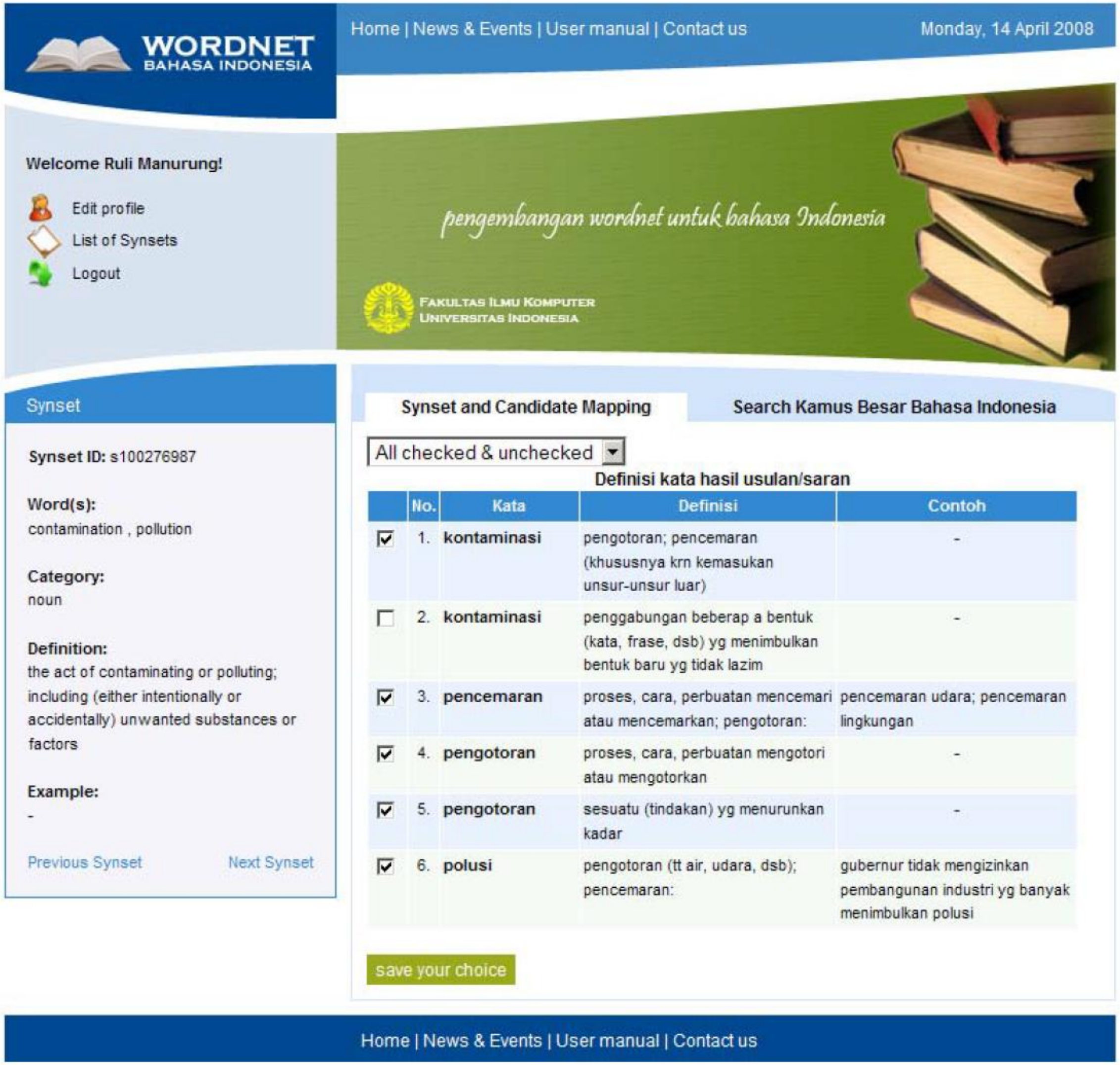
- Manually
  - Yes!
  - Time & Human Resource Limitation
- Solution
  - Crowdsourcing
  - Built the web application
    - Prepare the database
    - Develop the web application

Database Preparation

- Resource
  - PWN v3.0 Database (117.659 Synset & 147.306 Distinct Words)
  - KBBI Database (96.640 Unique Words Sense Definition)
  - Bilingual English-Indonesia Dictionary Database (For Suggestion in Web Application)


kbbi
kbbiid
lema
urut
kode
sublema
penggal
lafal
kelas
ragam
bidang
diabhs
sinon
rusil
definisi
contoh
makna
user
userid
username
password
nama
email
role
put in
usergroup
userid
groupid
makes
put in
map
userid
synsetid
kbbiid
date
put in
group
groupid
nama
has
put in
put in
kelas
kelasid
nama_kelas
pos_kelas
categorydef
categoryid
name
pos
refers to
synset
synsetid
categoryid
pos
definition
put in
groupsynset
groupid
synsetid
has
has
kelascategory
kelasid
categoryid
has
refers to
sample
synsetid
sampleid
sample
sense
wordid
casedwordid
synsetid
rank
lexid
tagcount
word
wordid
lemma
has

Web Application



Welcome Ruli Manurung!

Synset

Synset ID: s100276987

Word(s):
contamination , pollution

Category:
noun

Definition:
the act of contaminating or polluting; including (either intentionally or accidentally) unwanted substances or factors

Example:
-

Definisi kata hasil usulan/saran

	No.	Kata	Definisi	Contoh
☑	1.	**kontaminasi**	pengotoran; pencemaran (khususnya krn kemasukan unsur-unsur luar)	-
□	2.	**kontaminasi**	penggabungan beberap a bentuk (kata, frase, dsb) yg menimbulkan bentuk baru yg tidak lazim	-
☑	3.	**pencemaran**	proses, cara, perbuatan mencemari atau mencemarkan; pengotoran:	pencemaran udara; pencemaran lingkungan
☑	4.	**pengotoran**	proses, cara, perbuatan mengotori atau mengotorkan	-
☑	5.	**pengotoran**	sesuatu (tindakan) yg menurunkan kadar	-
☑	6.	**polusi**	pengotoran (tt air, udara, dsb); pencemaran:	gubernur tidak mengizinkan pembangunan industri yg banyak menimbulkan polusi

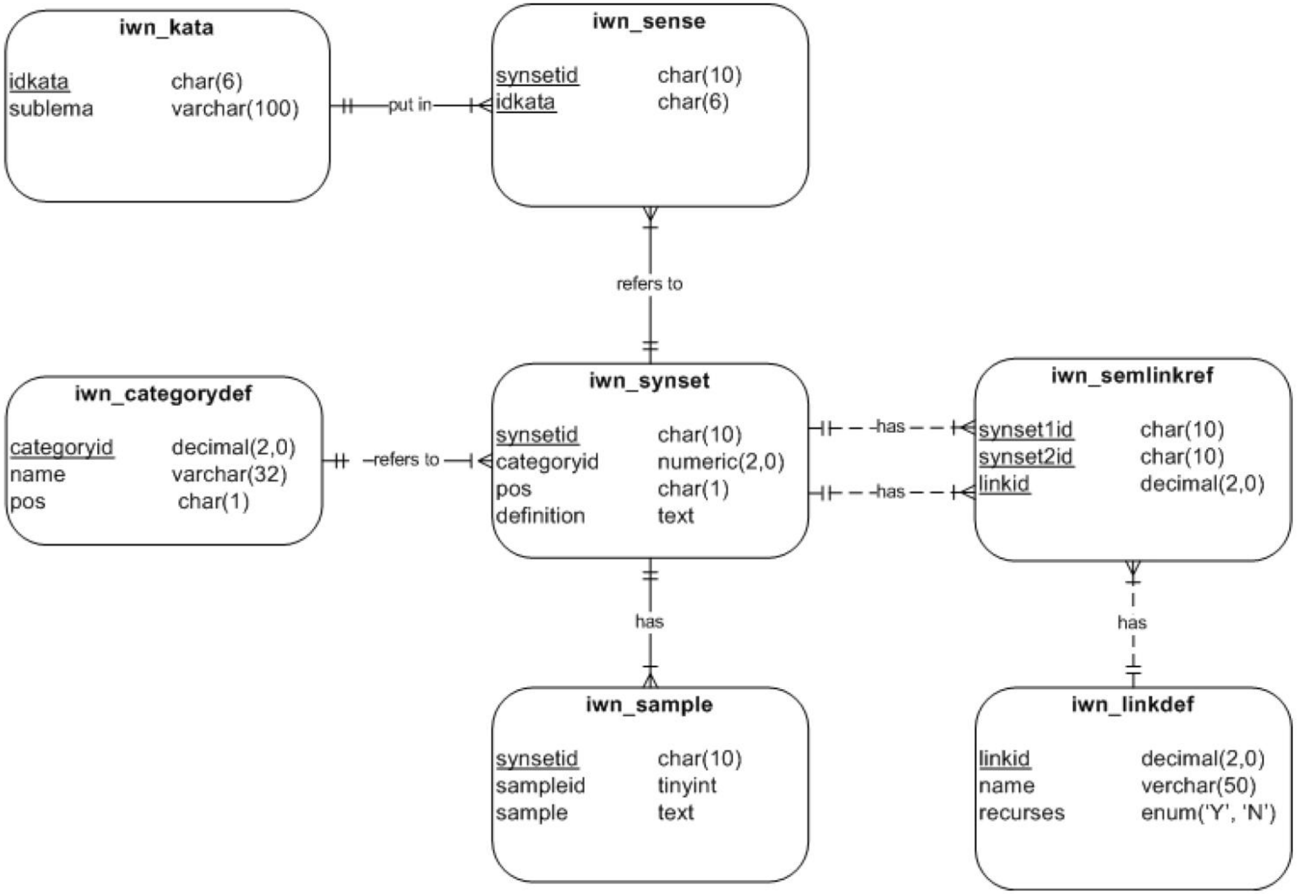
Obtaining Annotation

- 64 Annotators
- Independent judgment for every user
  - Prevent bias
- Map 3074 distinct sense definition to 1441 Unique Base Concept
- Ensure validity
  - Inter-annotator reliability using fleiss-kappa statistic

Building Indonesian Wordnet Database

- Last Step
  - Extract KBBI – PWN Mapping
- Process
  - Filter the mapping result using inter-annotator reliability
    - Set the degree of confidence
  - An IWN synset created from set of KBBI sense that are mapped to the same PWN synset
  - Copy the semantic relationship from PWN to IWN

IWN Database



iwn_kata	
idkata	char(6)
sublema	varchar(100)

iwn_sense	
synsetid	char(10)
idkata	char(6)

iwn_categorydef	
categoryid	decimal(2,0)
name	varchar(32)
pos	char(1)

iwn_synset	
synsetid	char(10)
categoryid	numeric(2,0)
pos	char(1)
definition	text

iwn_semlinkref	
synset1id	char(10)
synset2id	char(10)
linkid	decimal(2,0)

iwn_sample	
synsetid	char(10)
sampleid	tinyint
sample	text

iwn_linkdef	
linkid	decimal(2,0)
name	verchar(50)
recurses	enum('Y', 'N')