DataGuide w półstrukturalnych bazach danych

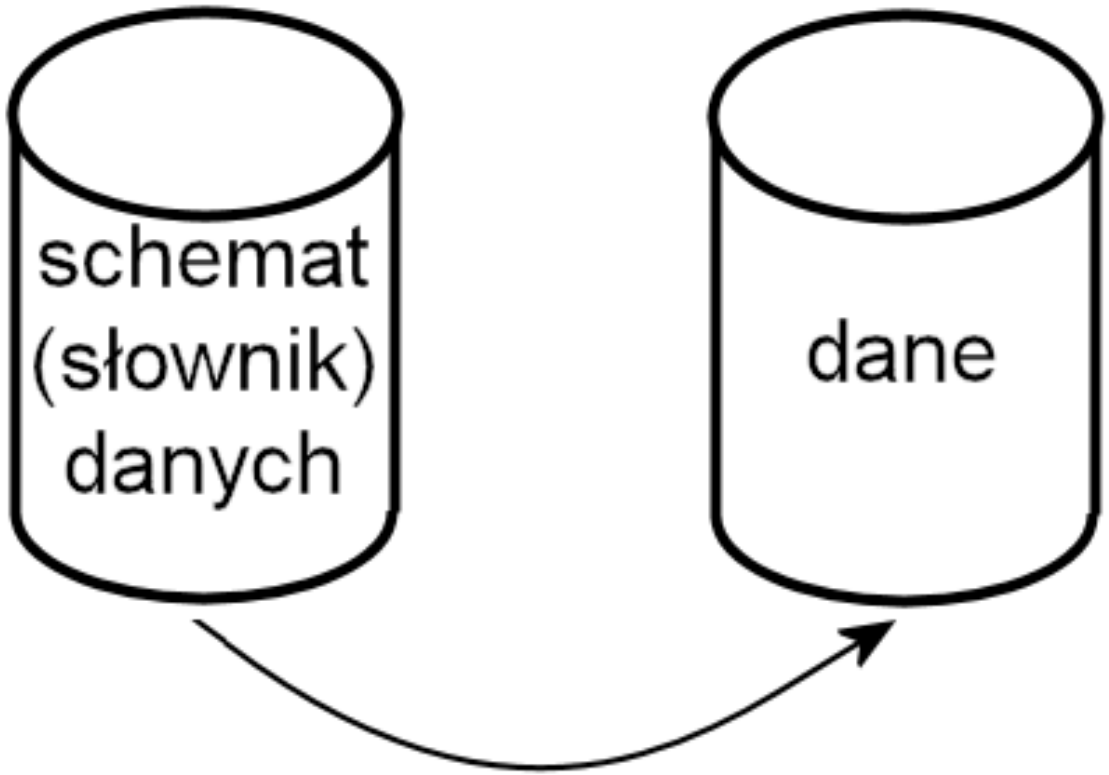
Marcin Jakubek

Plan prezentacji

- Schemat a dane
- Kilka słów o Lore
- DataGuide
- „Silny" DataGuide
- DataGuide a interakcja z użytkownikiem
- DG i optymalizacja zapytań
- Podsumowanie

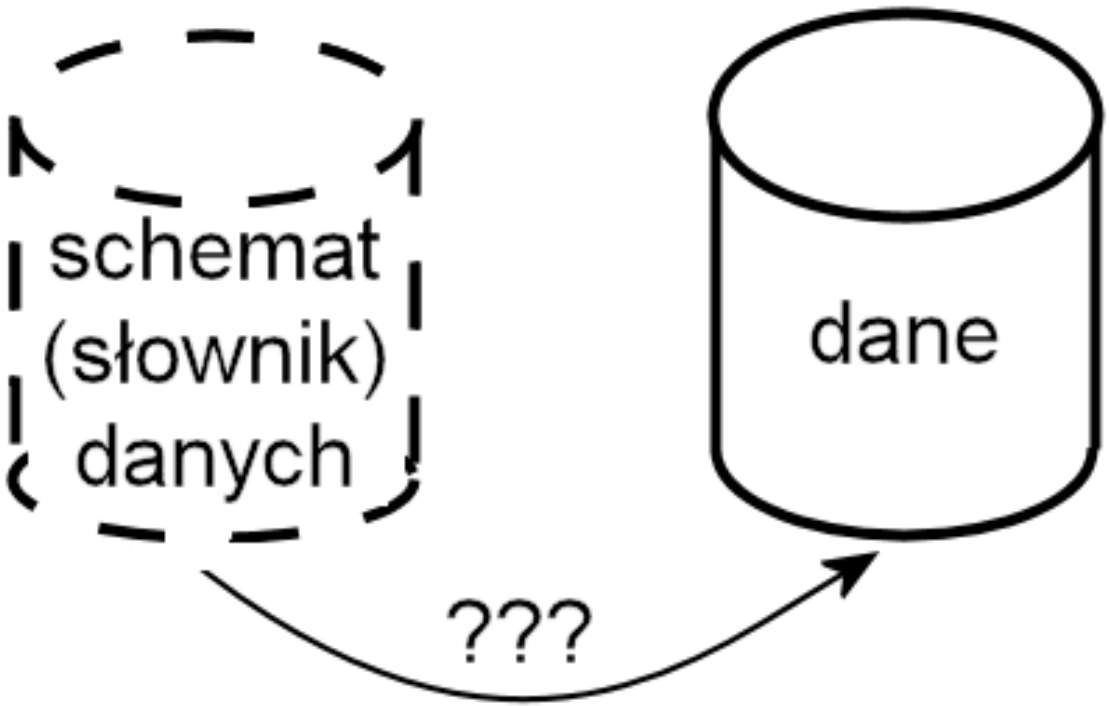
Schemat a dane

- Model ze strukturą (np. relacyjny)



Schemat a dane

- Model „pół"-strukturalny



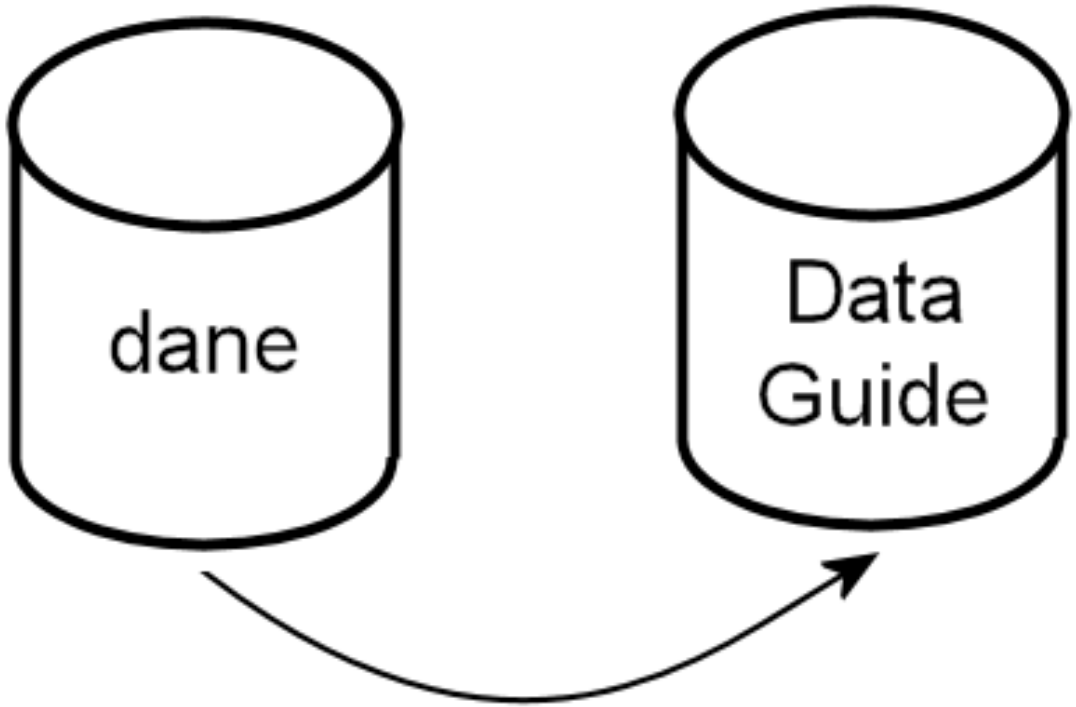
Schemat a dane

Po co nam schemat danych?

- dla użytkownika: informacja, co i gdzie jest w bazie, formułowanie zapytań
- dla SZBD: optymalizacja zapytań

Schemat a dane

DataGuide - „przewodnik po danych"



Lore - półstrukturalna BD

- Lore - *Light Object REpository*
- `http://infolab.stanford.edu/lore/`
- Model danych OEM
- Język zapytań Lorel

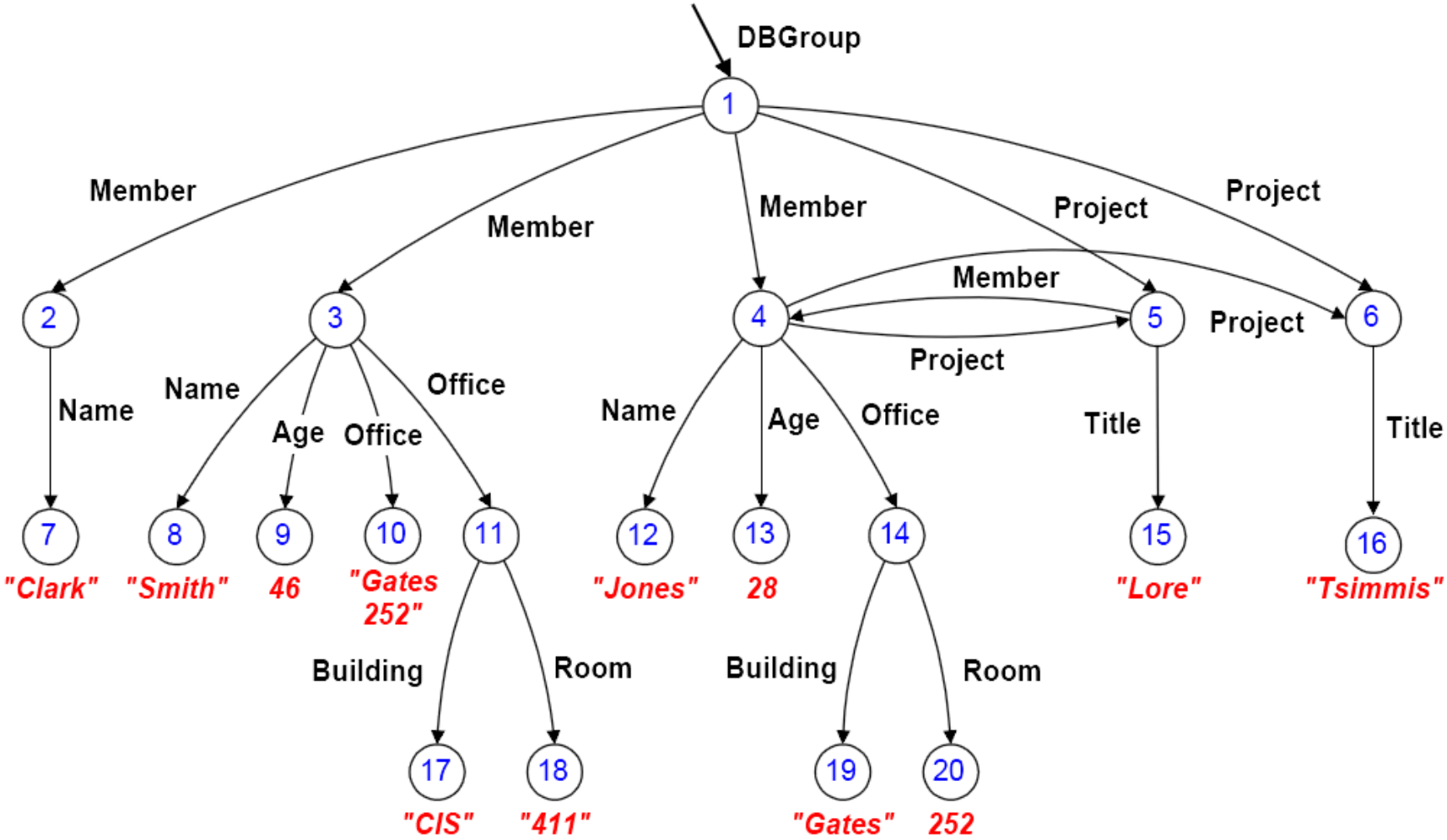
Lore - półstrukturalna BD

Model danych OEM - *Object Exchange Model* - to graf skierowany złożony z:

- obiektów (z oid):
  1. atomowych
  2. złożonych
  3. nazwanego obiektu korzeniowego
- etykietowanych krawędzi

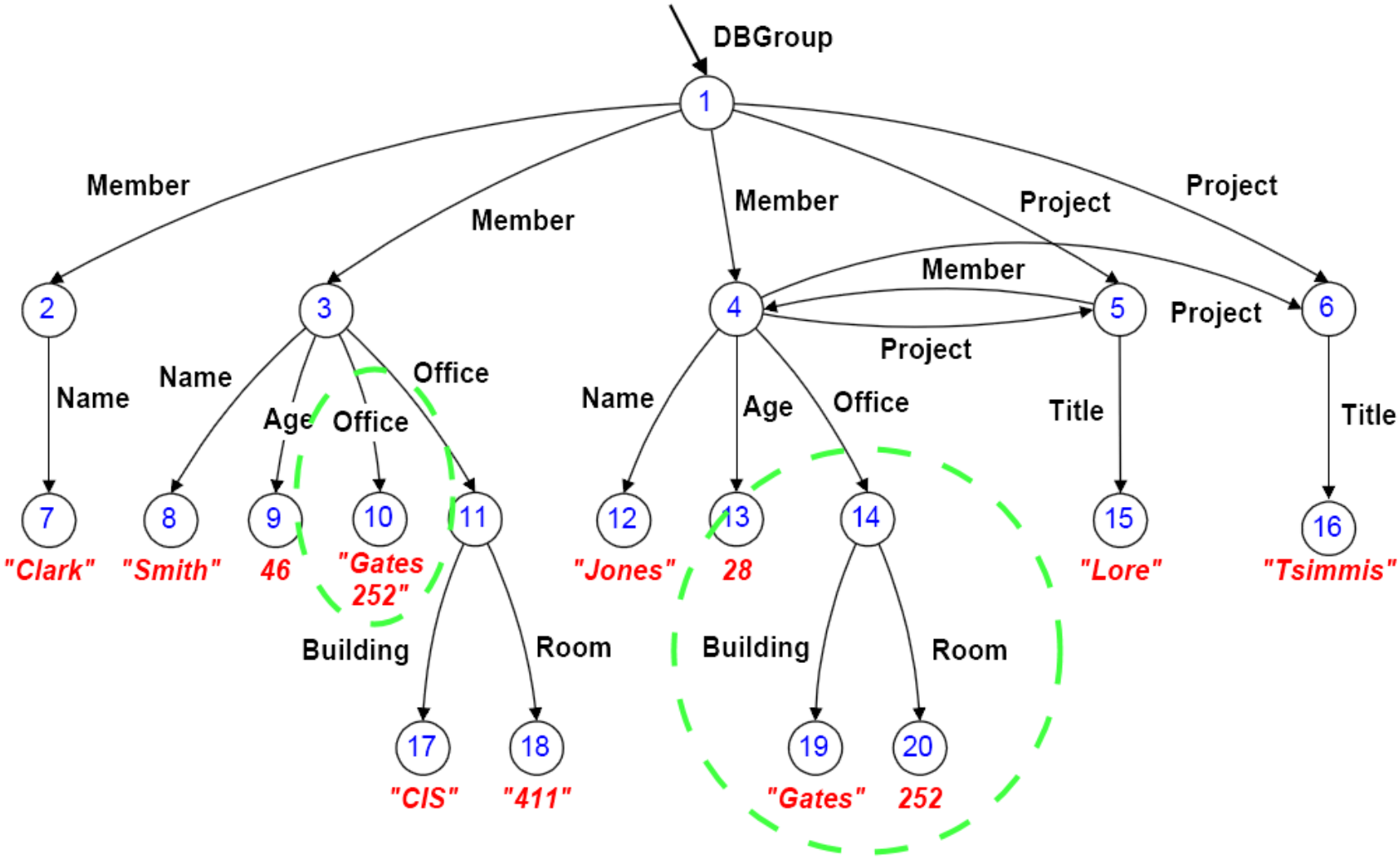
Lore - półstrukturalna BD

Przykładowy egzemplarz bazy Lore:



Lore - półstrukturalna BD

Przykładowy egzemplarz bazy Lore:



Lore - półstrukturalna BD

Język zapytań Lorel - *Lore language*

- „półstrukturalne" rozszerzenie Object Query Language
- zapytania ścieżkowe, np.:
  Select DBGroup.Member.Office where DBGroup.Member.Name = „Smith"
- możliwa opcjonalność, np.:
  select DBGroup.Member.Project where DBGroup.Member.#.(Office|Room) like “%252”
- zmienne ścieżkowe, etykietowe, kwantyfikatory, agregacja...

DataGuide

DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases

(1997)

Roy Goldman, Jennifer Widom

DataGuide

Def. 1. *Ścieżka etykiet* (label path) dla obiektu o to skończony ciąg etykiet $l_1.l_2.....l_n$ takich, że z obiektu o możemy przejść ścieżkę n krawędzi $(e_1...e_n)$, gdzie krawędź e_i ma etykietę l_i.
np.:
Office.Room dla obiektu 3
Member.Project.Member.Project dla obiektu 5.

DataGuide

Def. 2. *Ścieżka danych* (data path) dla obiektu o to naprzemienny ciąg etykiet i oid'ów $l_1.o_1.l_2.o_2.....l_n.o_n$ taki, że z obiektu o możemy przejść ścieżkę n krawędzi $(e_1...e_n)$ poprzez n obiektów $(x_1...x_n)$ gdzie krawędź e_i ma etykietę l_i a obiekt x_i oid o_i.

np.:

Member.2.Name.7 dla obiektu 1

Office.14.Room.20 dla obiektu 4.

DataGuide

Def. 3. Ścieżka danych d jest *egzemplarzem* ścieżki etykiet l jeśli ciąg etykiet w d jest taki sam jak w l.

np.:

Member.2.Name.7 dla **Member.Name**

Office.14.Room.20 dla **Office.Room**.

DataGuide

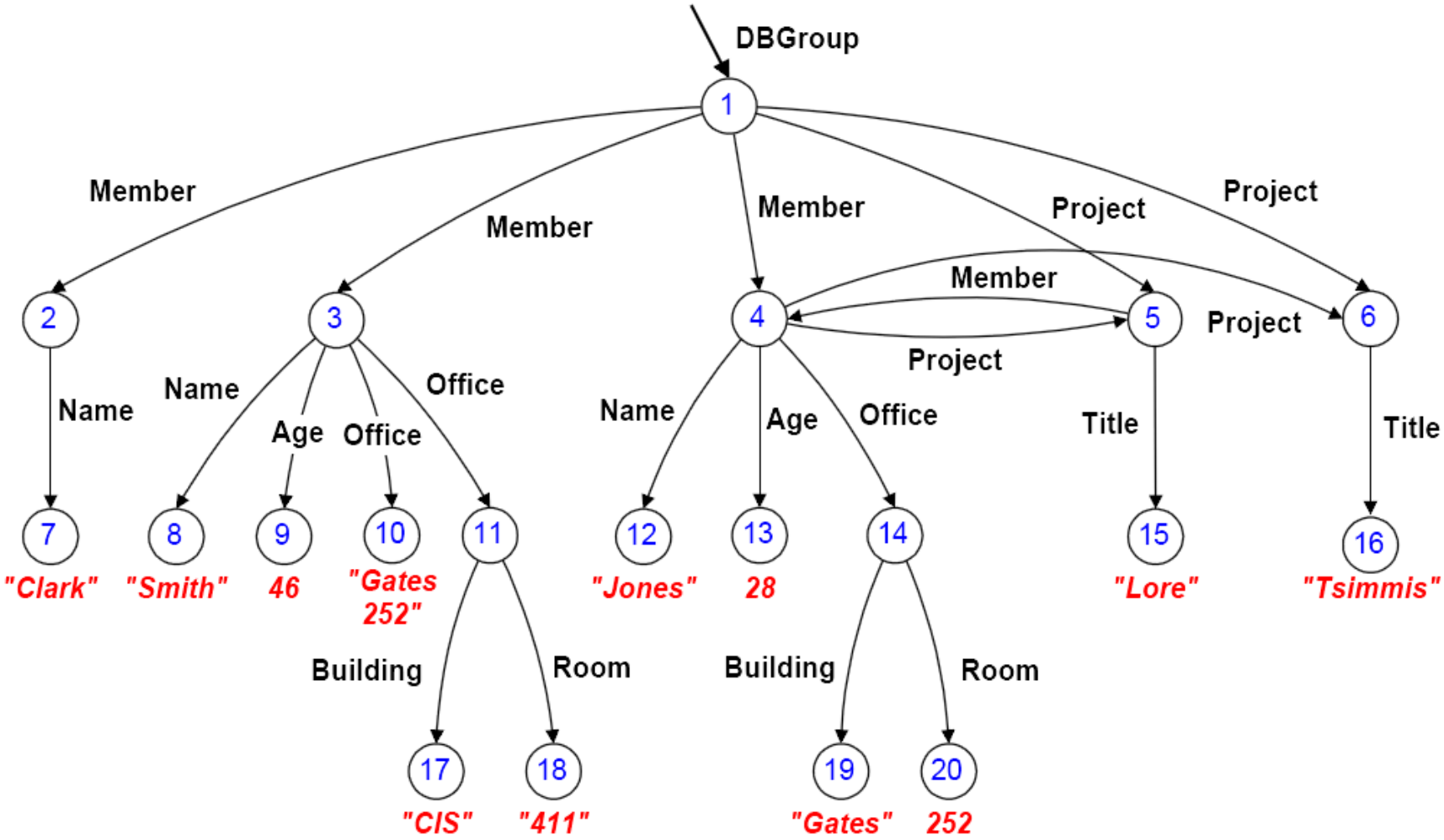
Def. 4. W bazie s, *zbiorem celów* (target set) nazywamy zbiór t oid'ów taki, że istnieje ścieżka etykiet l w s, dla której
$t = \{o | l_1.o_1...l_n.o$ jest egzemplarzem ścieżki $l\}$.
Tzn. t jest zbiorem wszystkich obiektów, które mogą być osiągnięte przez przechodzenie ścieżki l w bazie s. Oznaczenie: $t = T_s(l)$.
Wówczas mówimy, że poprzez ścieżkę l *osiągamy* każdy element zbioru t oraz że każdy element t *jest osiągalny* ścieżką l.
np.:
$\{5, 6\}$ jest zbiorem celów zarówno dla **Member.Project** i **Project.Member.Project**.

DataGuide

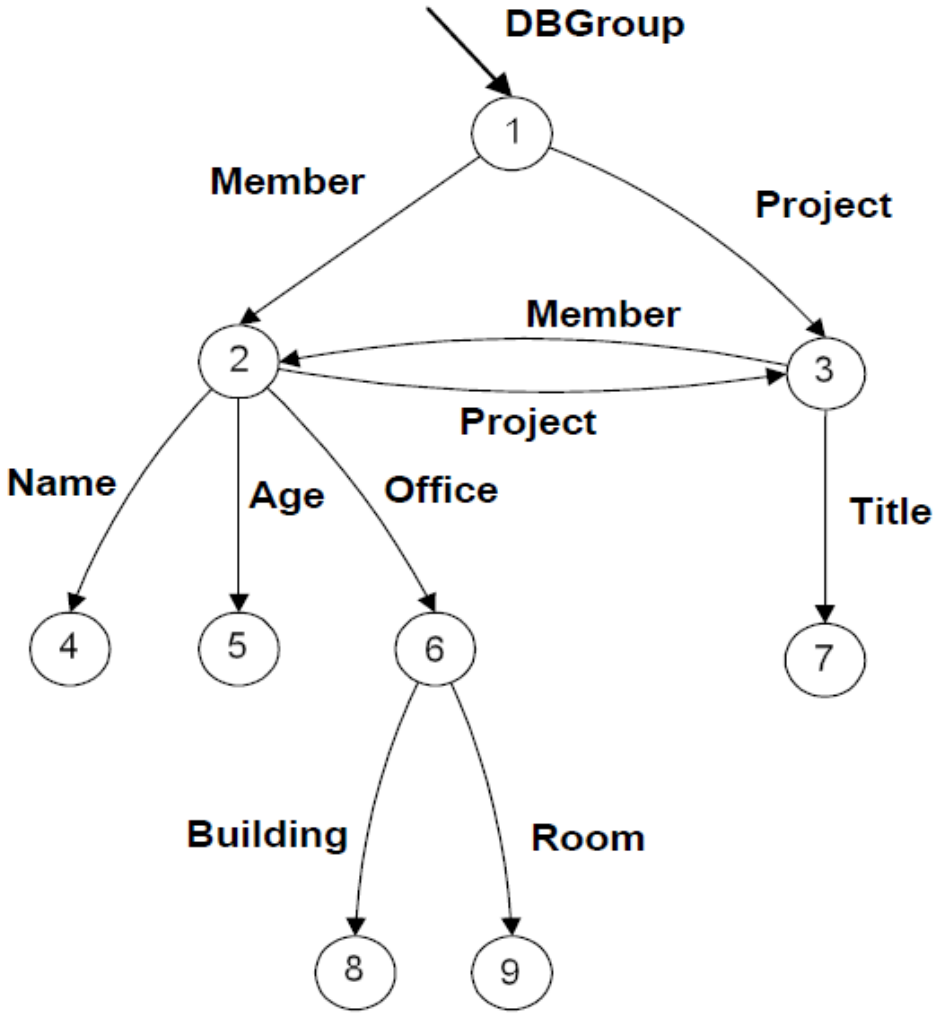
Def. 5. *DataGuide* dla bazy źródłowej s to obiekt OEM d taki, że każda ścieżka etykiet w s ma dokładnie jeden egzemplarz (ścieżkę danych) w d oraz każda ścieżka etykiet w d jest ścieżką etykiet w s.

Wniosek 1. Każdy zbiór celów w DataGuide jest singletonem.

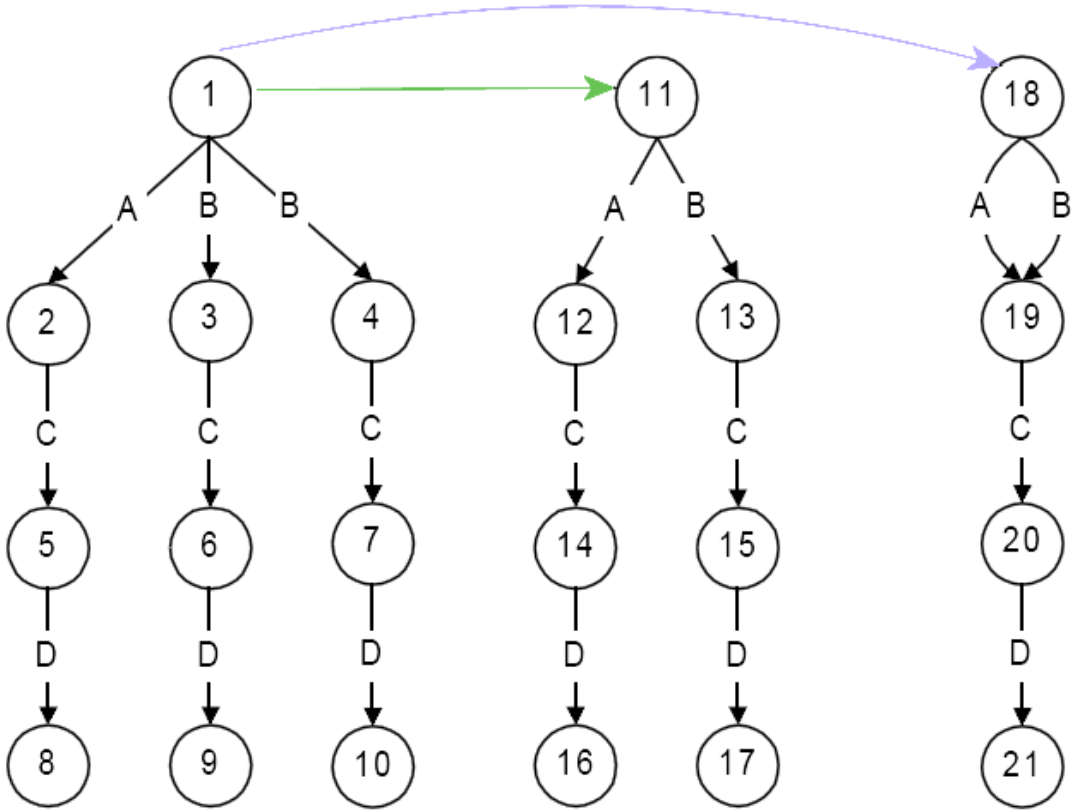
DataGuide - przykład



DataGuide - przykład



Wielokrotne DataGuide



Wielokrotne DataGuide

Dlaczego „większy" DataGuide jest „lepszy"?

1. Łatwiejsza aktualizacja
2. Możliwość przechowywania „adnotacji"

Wielokrotne DataGuide

Def. 6. W bazie s, dla danej ścieżki etykiet l, własność zbioru obiektów tworzącego zbiór celów ścieżki etykiet l w s nazywamy *adnotacją* (annotation) ścieżki l. Tzn. adnotacją ścieżki etykiet jest pewna charakterystyka zbioru osiągalnego w bazie przez tę ścieżkę.
np.:

- przykładowe dane osiagalne przez ścieżkę
- statystyki ścieżek wychodzących z obiektów w zbiorze celów ścieżki
- link do wszystkich obiektów w bazie osiągalnych przez ścieżkę

Strong DataGuide

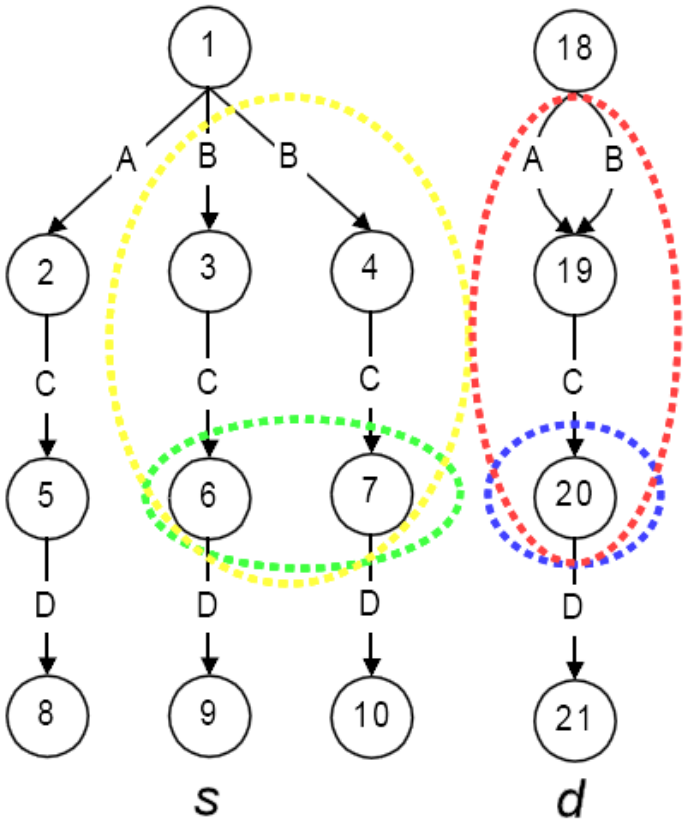
Kiedy adnotacje do zbiorów bazy mogą być przechowywane w obiektach w DataGuide?

Wtedy, gdy każdy zbiór ścieżek etykiet, które mają ten sam zbiór (singleton) celów w DataGuide, to zbiór ścieżek, które mają ten sam zbiór celów w bazie źródłowej.

Strong DataGuide

Def. 7. Niech d będzie obiektem DataGuide dla bazy s. Dla ścieżki l w s, niech $T_s(l)$ będzie zbiorem celów l w s, a $T_d(l)$ zbiorem (singletonem) celów l w d.
Niech $L_s(l) = \{m | T_s(m) = T_s(l)\}$, tzn. $L_s(l)$ jest zbiorem wszystkich ścieżek etykiet w s, które mają ten sam zbiór celów w s co l. Podobnie, niech $L_d(l) = \{m | T_d(m) = T_d(l)\}$, tzn. $L_d(l)$ jest zbiorem wszystkich ścieżek etykiet w d, które mają ten sam zbiór (singleton) celów w d co l.
Jeśli dla każdej ścieżki etykiet l w s $L_s(l) = L_d(l)$, to d nazywamy *silnym* (strong) DataGuide dla bazy s.

Strong DataGuide



$T_s(B.C)=\{6,7\}$

$T_d(B.C)=\{20\}$

$L_s(B.C)=\{B.C\}$

$L_d(B.C)=\{B.C, A.C\}$

$L_s(B.C) \neq L_d(B.C)$

Strong DataGuide

Tw. 1. Niech d będzie Strong DataGuide dla bazy s. Jeśli adnotacja p pewnej ścieżki etykiet jest przechowywana w obiekcie o osiągalnym przez l w d, to p opisuje zbiór celów w s każdej ścieżki etykiet (w s), która osiąga o w d.

Tw. 2. Niech d będzie Strong DataGuide dla s. Każdy zbiór celów t w s jest z definicji zbiorem celów pewnej ścieżki etykiet l. Niech F będzie przyporządkowaniem $l \longrightarrow T_d(l)$ - czyli singletonu w d. Wówczas F wprowadza odpowiedniość $1-1$ między zbiorami celów w bazie s a obiektami (singletonami) w DataGuide d.

Strong DataGuide

Zalety Strong DataGuide:

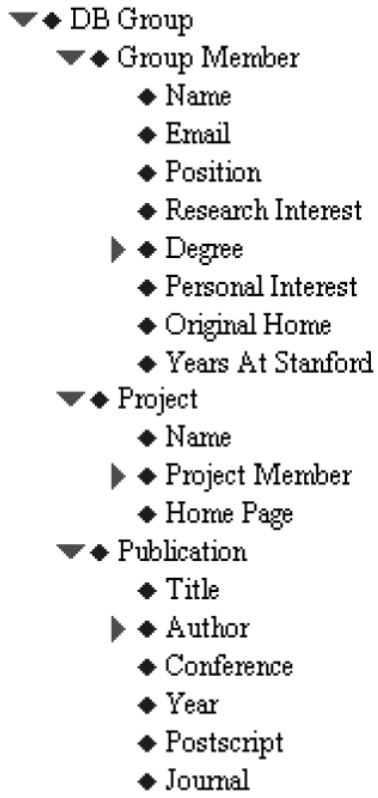
- odpowiedniość 1-1 między zbiorami celów ścieżek w bazie a obiektami w DataGuide
- łatwa rozbudowa
- dla drzewiastych baz czas budowy liniowy od rozmiaru bazy

Wady (dla baz z cyklami):

- nawet wykładniczy czas budowy
- Strong DataGuide może być nawet wykładniczo większy od bazy

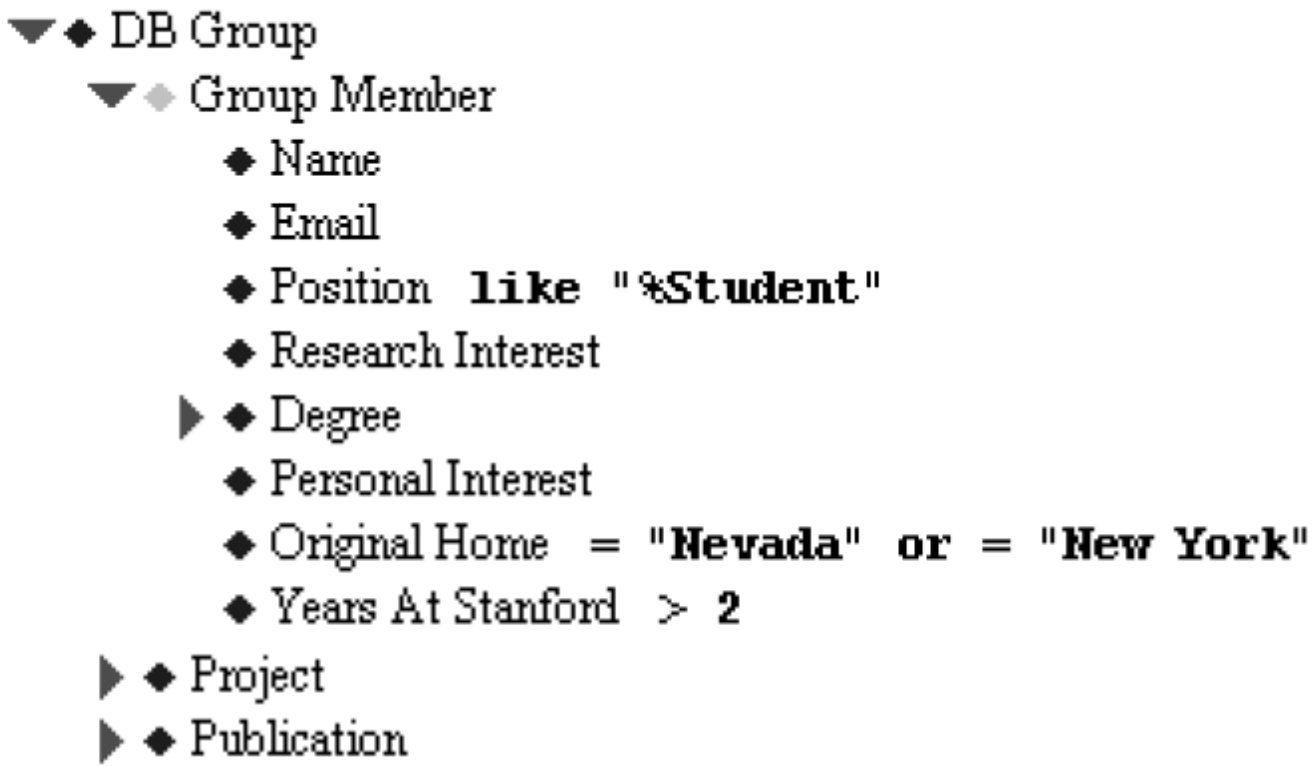
DataGuide dla użytkownika

DataGuide - zwięzłe, aktualne „podsumowanie" struktury bazy do zastosowania w GUI:



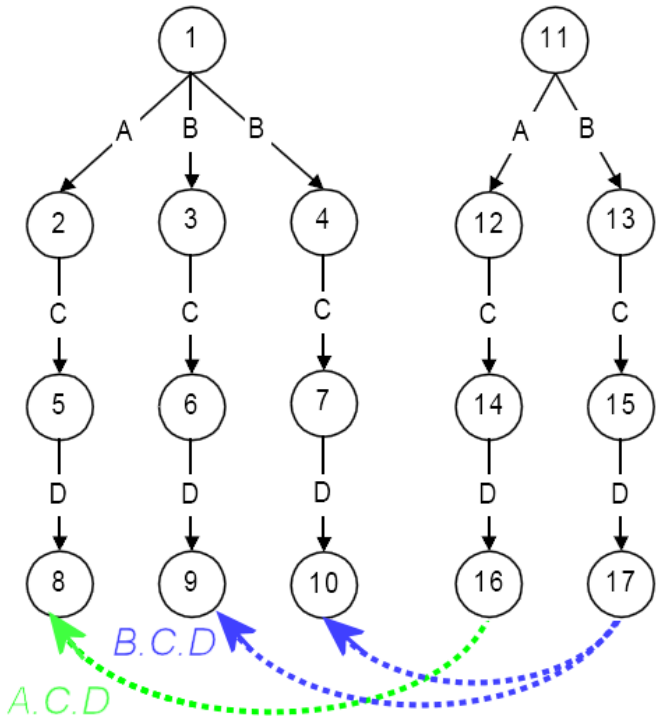
DataGuide dla użytkownika

oraz przy formułowaniu zapytań:



DataGuide dla SZDB

W Strong DataGuide obiekty mogą przechowywać linki do zbiorów celów osiągalnych w źródłowej bazie przez dane ścieżki etykiet:



DataGuide dla SZDB

Przykład dla bazy Stanford DBGroup:
Select DBGroup.Member.Publication.Troff
Założenia.:
10 000 DBGroup.Member, średnio po 100 publikacji, 1 publikacja typu **Troff**
⇓
Bez DataGuide - 1 000 000 obiektów do przejrzenia.

Z DataGuide - 6 obiektów: korzeń DataGuide→ DBGroup→ Member→ Publication→ Troff → link do obiektu w bazie

Podsumowanie

DataGuide - „schemat generowany z danych":

- daje pogląd na strukturę danych (GUI, zapytania)
- szybki dostęp do obiektów osiągalnych przez dane ścieżki (optymalizacja)
- szybka weryfikacja istnienia ścieżki
- dla zacyklonych baz danych - nawet wykładnicza budowa względem rozmiarów źródłowej bazy

Dziękuję za uwagę!