YARN: НАЧАЛО / YARN BEGINS

П. И. Браславский
М. Ю. Мухин
О. Н. Ляшевская
А. А. Бонч-Осмоловская
А. А. Крижановский
П. В. Егоров

Диалог-2013

WordNet для русского языка

- Необходимый ресурс для автоматической обработки текстов (наряду с морфологическим анализатором, большим аннотированным корпусом и т.п.).

- Общедоступный большой ворднет на материале русского языка пока не существует!

Анонсированные ресурсы

- Оригинальные лексические базы:
  - RussNet (рук. И. В. Азарова)
  - Тезаурус РуТез (НИВЦ МГУ).
- Проекты, связанные с переводом и адаптацией PWN:
  - СПб, Госуниверситет путей сообщения [Сухоногов, Яблонский 2005]
  - Новосибирский ГУ [Гельфенбейн и др.]

- степень разработки?
- незавершенность или закрытость

Пути решения организационных проблем

- Использование существующих источников (словарей, корпусов)
- Автоматические методы
- Crowdsourcing

Yet Another RussNet (YARN)

- Инициативная группа – представители УрФУ, Kontur Labs, ИММ УрО РАН, ВШЭ
- участники из Санкт-Петербурга, Казани, Челябинска, Томска, Петрозаводска.

- http://groups.google.com/group/yarn_org/

Основная идея YARN

- Сочетание традиционных принципов создания ворднетов и wiki-подхода.
- Открытость, открытость и еще раз открытость.

Wiki-принцип

- Большое количество участников (например, студенты-лингвисты).
- Централизованное хранение данных, редактирование тезауруса через веб-интерфейс.
- Контроль процесса редактирования и качества.

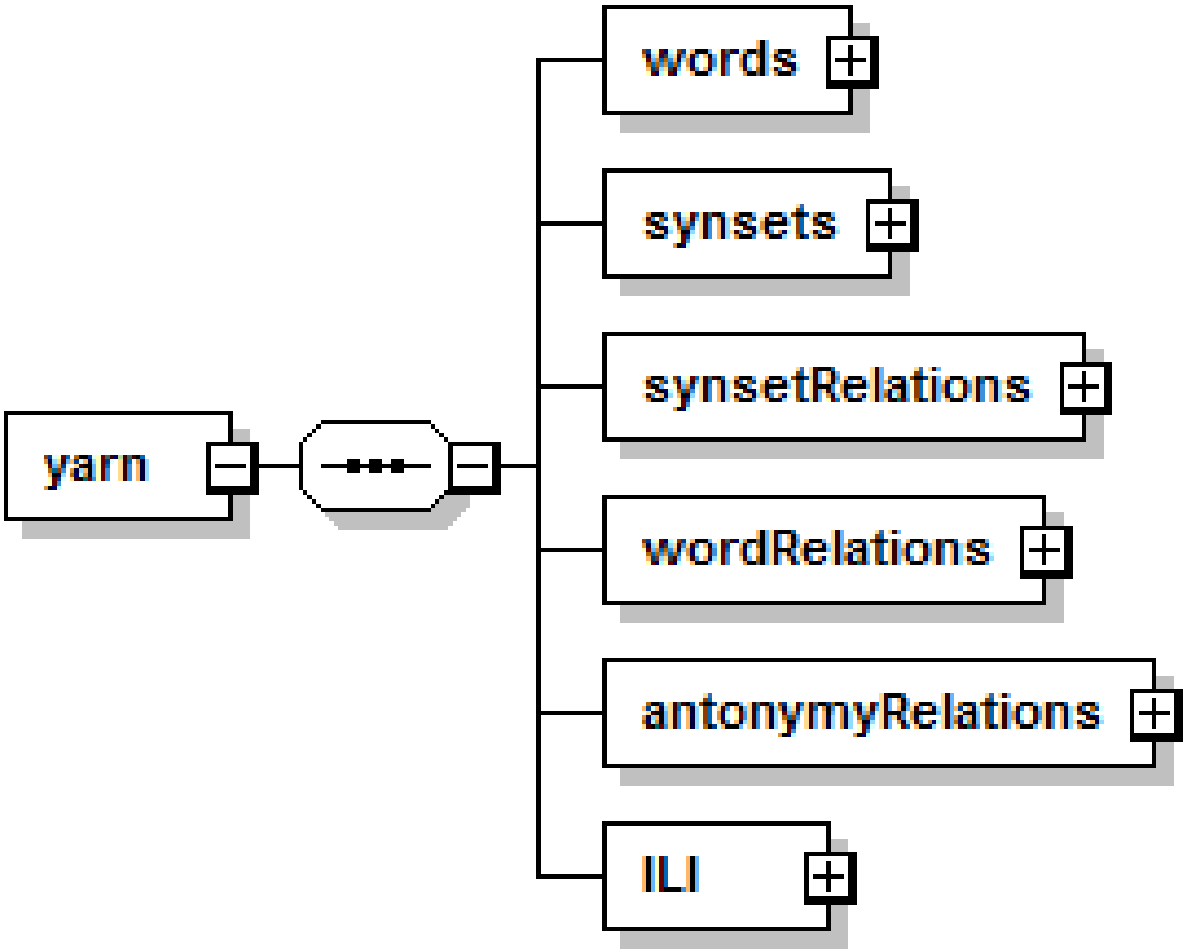
Лингвистические принципы YARN

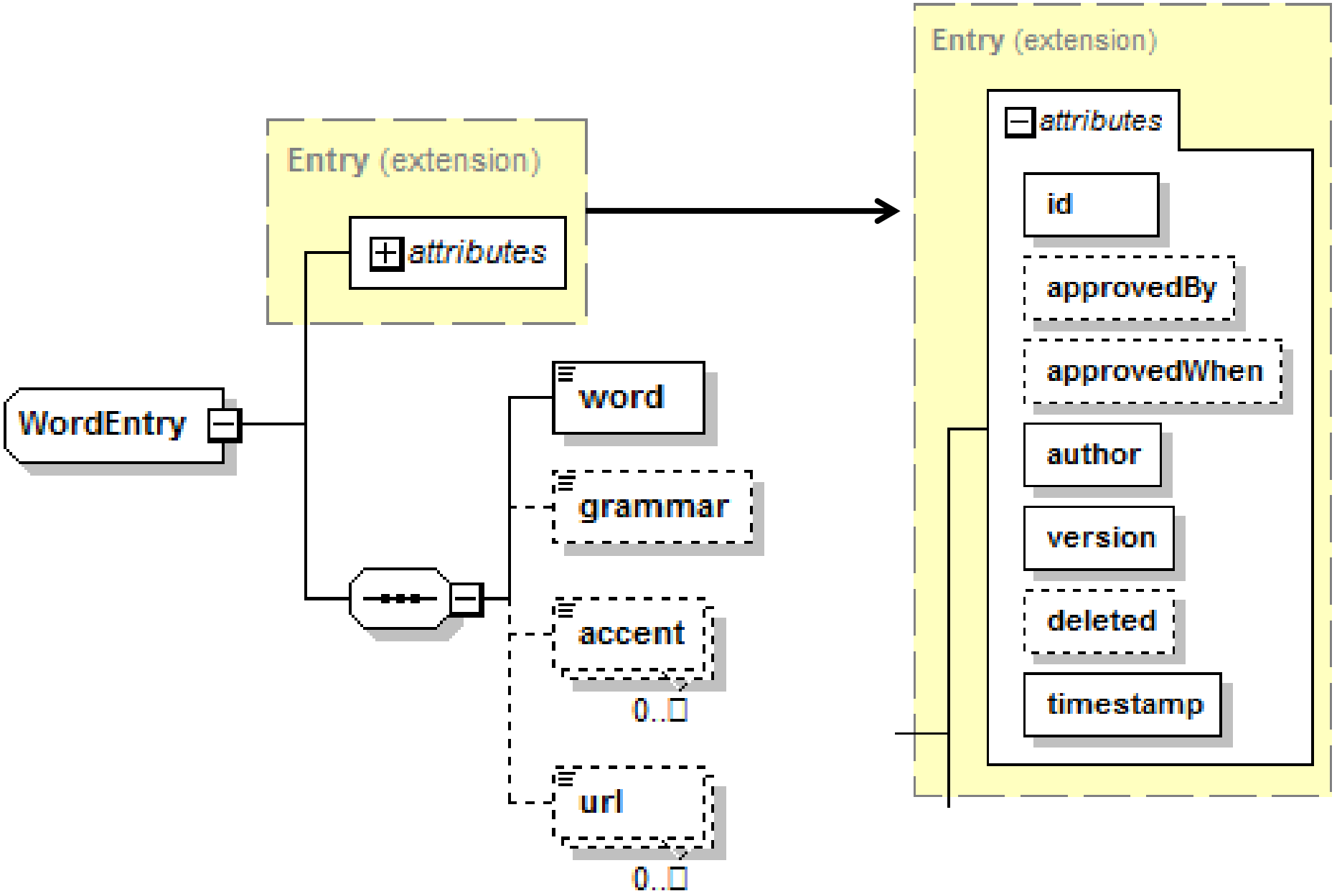
- YARN ориентируется на мировую практику создания wordnet-подобных ресурсов.
- Начальное лексическое наполнение – существительные, прилагательные и глаголы.

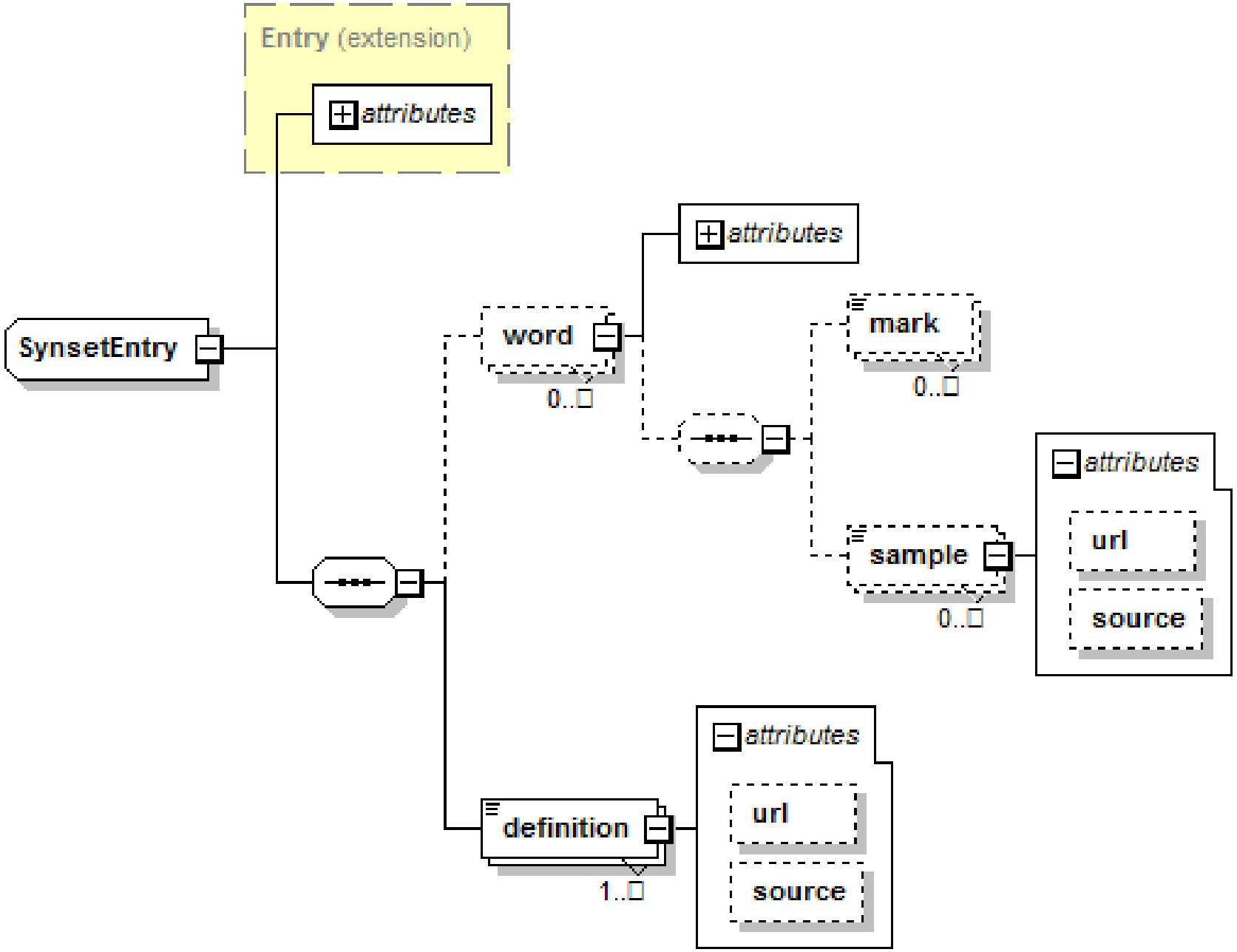
Стандартная WN-подобная организация YARN

- Основа – синсеты (группы синонимов и квазисинонимов).
- Синсеты упорядочены между собой иерархически и связаны отношениями гиперо/гипонимии, антонимии, холонимии/меронимии

Формат XML: компоненты схемы




Entry (extension)
attributes
WordEntry
word
grammar
accent
0..□
url
0..□
Entry (extension)
attributes
id
approvedBy
approvedWhen
author
version
deleted
timestamp


Entry (extension)
attributes
SynsetEntry
word
0..∞
attributes
mark
0..∞
attributes
url
source
sample
0..∞
definition
1..∞
attributes
url
source

Первые шаги

- 20 000 частотных существительных (НКРЯ)
- Данные русского Викисловаря
- Сборка синсетов

Приглашаем к сотрудничеству

- http://russianword.net
- http://groups.google.com/group/yarn_org/

YARN — проект в стадии становления

- максимальное расширение числа участников редактирования словарных данных
- привлечение возможностей автоматизированного извлечения лексикографической информации
- открытость проекта в плане его дальнейшего развития и использования.

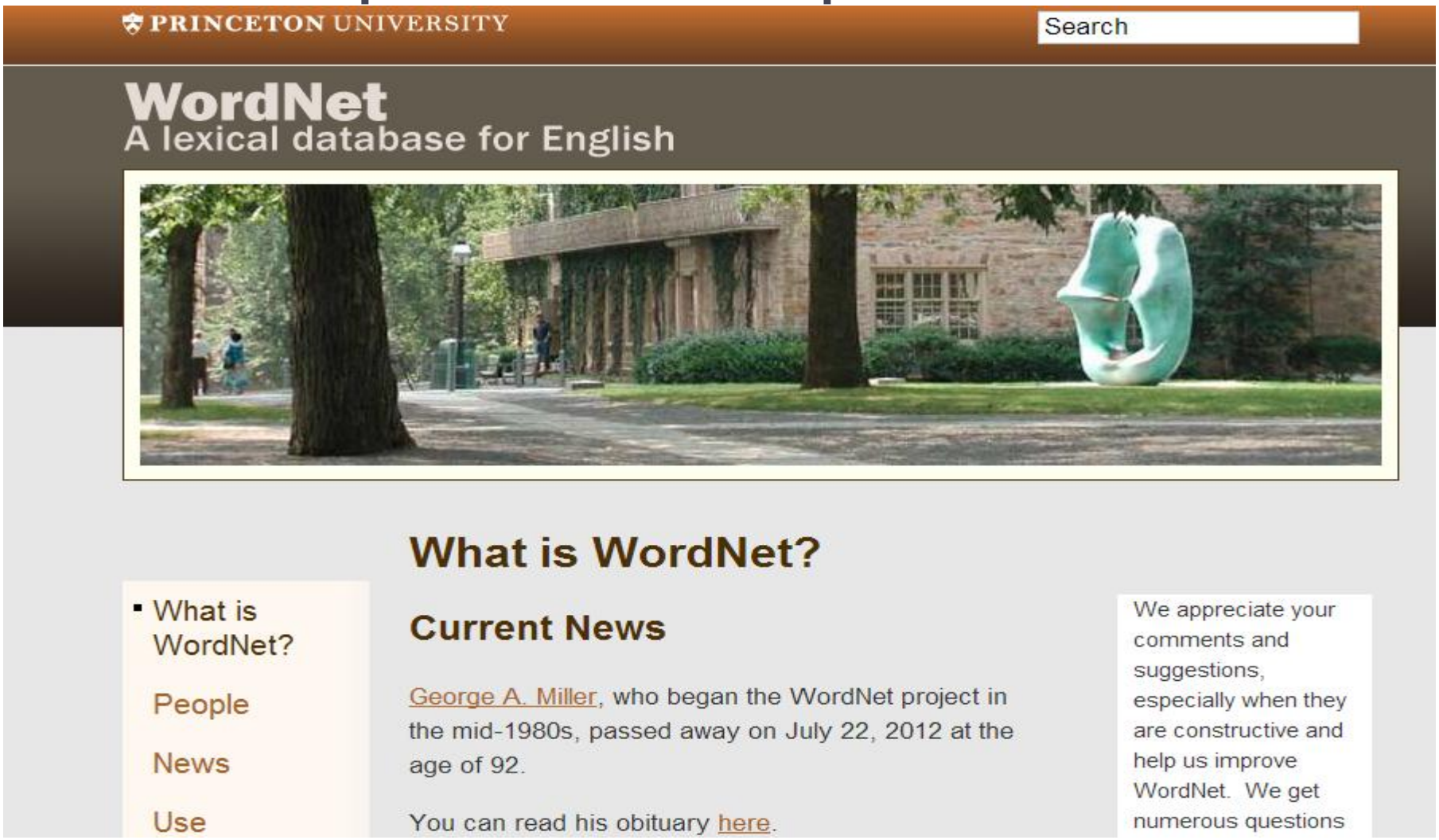
Благодарности

- Исследование осуществляется при финансовой поддержке РГНФ (проект № 13-04-12020 «Новый открытый электронный тезаурус русского языка»).
- Благодарим участников группы *yarn_org* за активность, замечания и предложения.

Спасибо за внимание!

Браславский П.И.	pb@kontur.ru
Мухин М.Ю.	mfly@sky.ru
Ляшевская О.Н.	olesar@gmail.com
Бонч-Осмоловская А.А.	abonch@gmail.com
Крижановский А.А.	andrew.krizhanovsky@gmail.com
Егоров П.В.	pe@kontur.ru

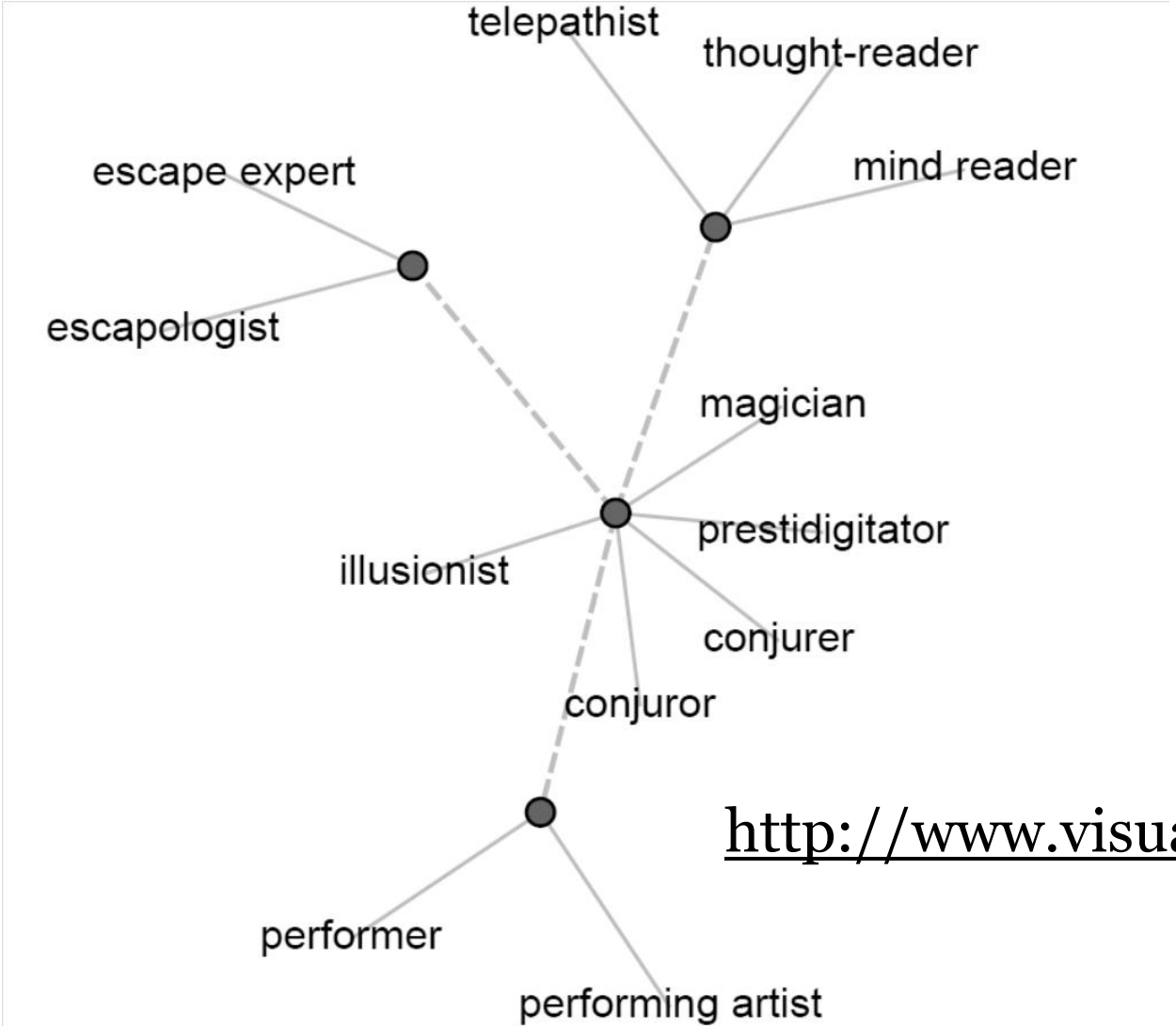
P.S. Princeton Wordnet, started in 1986. http://wordnet.princeton.edu/



Princeton Wordnet (PWN)

- Стандартный подход к организации таких ресурсов.
- Сегодня *ворднетами* называют любые лексические базы, построенные по схожим принципам: они состоят из синсетов (*synset*, от *synonym set*) – «смыслов», которые выражаются набором квазисинонимов. В свою очередь синсеты связаны между собой различными отношениями: гипероним/гипоним, мероним/холоним и др.

Наглядное представление фрагмента PWN



http://www.visualthesaurus.com/