

ВИКИ-КОНФЕРЕНЦИЯ 2007

ТЕЗИСЫ ДОКЛАДОВ

Санкт-Петербург 2007

Содержание

Вступительное слово.....	3
Будущее Википедии.....	4
Проблемы малых разделов.....	5
Проблемы развития чувашского раздела.....	7
Достоверность информации в Википедии. Авторитетность источников информации и проект «Выверка статей».....	8
Оценка результатов поиска семантически близких слов в Википедии.....	9
Связность статей русской Википедии.....	10
Проблемы хронологического упорядочения статей и данных.....	12
Социально-политические науки в Википедии.....	13
Перекрёстная каталогизация в Википедии.....	14
Деление информации в Википедии по темам: способы размежевания и викисоциальные проблемы.....	15
Википедия и проблемы.....	16
Многоязычная Википедия: языковой помощник для физика и научно-технического переводчика.....	17
О работе арбитражного комитета русского языкового раздела Википедии.....	20
Брокгауз и Ефрон — вики-технологии XIX века.....	20

Вступительное слово

Павел Каганер

«Википедия» (<http://wikipedia.org>) — это свободная энциклопедия в Интернете. В отличие от других проектов, она создаётся не государственными органами и не коммерческими организациями, а совместными усилиями многих тысяч добровольцев.

В Википедии нет сложных процедур. Любой может внести в статьи полезные правки. Ваши изменения сразу становятся доступными всему Интернету. Такая простота является основой популярности Википедии. Недобросовестные изменения быстро нейтрализуются благодаря продуманной технической и социальной организации проекта.

Википедия состоит из более 250 языковых разделов. Раздел на русском языке занимает 11-е место по количеству статей. 4 сентября 2007 г. количество статей в Википедии на русском языке превысило 200-тысячный рубеж и продолжает стремительно увеличиваться.

За семь лет развития русская Википедия стала абсолютным лидером среди всех русскоязычных энциклопедий Интернета по количеству статей. Она более чем в два раза опережает Большую советскую энциклопедию и более чем в полтора — Энциклопедический словарь Брокгауза и Ефрона.

Индекс цитирования русской Википедии, измеренный «Яндексом», составляет 12 000. Сайт Википедии находится на 38-42 месте среди всех сайтов Рунета. 29 ноября 2006 года проект получил государственную награду «Премия Рунета» в номинации «Наука и образование». В начале 2007 года при присуждении премии РОТОР в номинации «Научно-образовательный сайт года» русская Википедия заняла второе место, уступив лишь старейшему российскому научно-популярному журналу «Вокруг света». В номинации «Интернет-сообщество года» раздел на русском языке также занял второе место.

Википедия превосходит бумажные энциклопедии огромным количеством иллюстраций. Мультимедийные возможности электронной энциклопедии позволяют не только проиллюстрировать статью, но и озвучить. Большая часть иллюстраций используется совместно всеми разделами Википедии, и таких иллюстраций уже больше полутора миллионов.

Тексты и изображения проверяются на отсутствие нарушений авторских прав. Всё текстовое содержимое Википедии распространяется исключительно на условиях свободной лицензии GNU FDL, и большинство её иллюстраций распространяются по свободным лицензиям.

География участников русского раздела охватывает, помимо России, страны ближнего и дальнего зарубежья, включая Австралию, Белоруссию, Бельгию, Германию, Данию, Израиль, Латвию, Литву, США, Украину, Эстонию, Южную Корею и многие другие. Зарегистрировано более 78 тысяч учётных записей.

Участники Википедий на разных языках встречаются очно на локальных викивстречах и международных встречах «Викимания». Являясь одновременно научной конференцией и социальным событием, «Викимания» собирает вместе участников различных проектов фонда «Викимедиа» для обмена идеями, построения новых партнёрств, докладов об исследованиях и результатах проектов. Последняя «Викимания» состоялась 3—5 августа 2007 в городе Тайбэй на Тайване.

Будущее Википедии

Ю. Л. Колесников, А. Н. Кондратьев, Г. Е. Скворцов

Википедия, с одной стороны, – специфический феномен эпохи зарождения компьютеров и Интернета. С другой стороны, Википедия сохраняет черты средневековых энциклопедий: собрание известных знаний в форме статей, посвящённых отдельным темам. **Противоречие** между традиционной организацией знаний и информационными инновациями частично решила технология вики. Она предоставила возможность оперативного обновления энциклопедии людьми, находящимися на разных континентах.

Сейчас Википедия интенсивно развивается. Проблемы современного этапа (некомпетентность первых энтузиастов, «вандализм», войны откатов, перекосы в тематике и др.) находят своё решение благодаря сложившемуся феномену коллективного разума.

На S-образной кривой развития Википедия находится на переломе от первого этапа медленного становления ко второму этапу интенсивного роста. Кроме экстенсивных показателей развития (рост количества статей, увеличение количества участников, приход в Википедию именитых учёных, улучшение интерфейса, ускорение серверов) неизбежно произойдут **качественные изменения**. Важно заранее осознать, что в настоящем виде Википедия будет не вечно, что наряду с мелкими проблемами текущего этапа будут появляться новые большие проблемы. И надо быть готовыми к ним и предопределять пути плодотворного решения.

Одним из качественных изменений, на наш взгляд, должна стать трансформация Википедии в единый банк системы человеческих знаний, обладающий новыми принципами структурирования. Нынешнее представление Википедии в виде отдельных статей не отвечает системному статусу и существенно снижает продуктивность, то есть возможность эффективного использования этого гигантского набора знаний.

Задачей номер один для повышения плодотворности, по нашему мнению, является превращение набора знаний в систему, создание качественно новых принципов **организации знаний**. На начальном этапе организация знаний, по существу, сводится к систематизации. Опыт систематизации в ряде наук с большим объёмом информации (например, биологии, геологии, социологии) указывают способы действия. В то же время, несогласованность развития отдельных отраслей наук приведёт к появлению противоречий при создании единого представления всех знаний. Неизбежно должны появиться единые принципы организации, развития и представления каждой из отдельных областей, что окажется мощным толчком к прогрессу наук, находящихся на первых этапах развития.

Википедия может оказаться интегратором как учёных, так и знаний, а также стать зачинателем информационной эры систематизации научных представлений. Будет найден механизм эффективного взаимодействия всех научно-исследовательских подразделений мира. Концептуальные преобразования представления знаний в Википедии неизбежно приведут к появлению новых принципов социальной организации и технического обеспечения Википедии. Произойти это должно на новом этапе развития Википедии, который, судя по темпам развития, должен произойти довольно скоро.

Проблемы малых разделов

Вячеслав Иванов, Николай Плотников, Денис Сахарных

Доклад содержит краткий обзор основных проблем малых разделов (прежде всего на языках народов Российской Федерации, которые мы условно называем по тексту «малыми языками») и некоторые предложения о том, как сообщество Википедии на русском языке могло бы участвовать в разрешении этих проблем.

Обоснование важности

Существование разделов на малых языках способствует расширению социального функционирования этих языков, выводит малые языки в их письменной форме в массовое употребление, повышает престиж малых языков и, в конечном итоге, способствует их сохранению и развитию.

Для большинства языков Российской Федерации подготовка универсальной энциклопедии иными средствами не представляется возможной.

Исходя из ценности языкового разнообразия, заявленной в документах ЮНЕСКО, а также в законодательстве РФ и отдельных российских регионов, помощь малым разделам Википедии кажется нам полезным и благородным делом.

Проблемы малых разделов

На этапе создания

Если первые языковые разделы на языках народов РФ создавались без какой-либо заявки, по алфавиту (ab:, av:, ce:, cv: и другие), то впоследствии, уже к началу 2005 года, требования фонда Викимедиа ужесточились. Для создания осетинского, удмуртского и лакского разделов потребовалось заполнение заявки (на английском языке) и отстаивание этой заявки — не для всех языков находятся энтузиасты, готовые этим заниматься. В последнее время для создания нового раздела требуется успешное его ведение в инкубаторе (на этой стадии сейчас находятся разделы на ингушском, якутском и некоторых других языках).

Возможная помощь вики-сообщества здесь может сводиться к отслеживанию правильности заполнения заявки и к технической помощи (викификация, создание шаблонов и т. п.) в развитии раздела в инкубаторе.

В уже существующем разделе

С существующих разделах практически без исключения основной проблемой является низкое качество содержимого и неразвитость сообщества (малое число редакторов вообще, отсутствие или недостаточное число постоянных редакторов).

Привлечь редакторов может повышение внешней привлекательности проекта — оформление главной страницы, создание шаблонов «как в больших Википедиях», реализация удобной навигации и другие действия, доступные опытным википедийцам и без знания языка раздела.

Массовая заливка однородного содержимого

Массовая роботизированная заливка содержимого в малые разделы несколько дискредитирована в последнее время инцидентами, когда такая заливка осуществлялась ради привлечения внимания к языку (в случае с разделом на волапуке, например). Однако разумное применение роботов для создания востребованных начатков статей очень нужно для обустройства любого малого раздела.

При помощи роботов давно и успешно создаются начатки однотипных статей о годах, городах и другие.

Поиск и сканирование текстов старых книг

На многих языках РФ в 1920-50 годах издавались книги в том числе справочного содержания, на которые истёк срок копирайта. Некоторые из этих текстов не утратили своей актуальности (как исследования по фольклору или школьные учебники по «неустаревшим» предметам). Поиск и подготовка таких текстов к дальнейшей переработке может быть серьёзным подспорьем в развитии малого раздела.

Викификация статей

Зачастую не нужно знания языка, чтобы помочь в викификации статьи — замене дефисов на тире, оформлении внутренних ссылок и заголовочного слова статьи, приведении в читабельный вид таблиц и т. д.

Тексты в малых разделах создают в том числе новички, плохо представляющие себе язык вики-разметки и другие особенности проекта. В отличие от русского раздела, не всегда есть активное сообщество, готовое сразу викифицировать вклад новичка.

Участие в обсуждениях

Нередко в национальных разделах возникают обсуждения, в которых участие стороннего медиатора или эксперта весьма желательно. Если такие обсуждения ведутся на русском (что бывает нередко), появление одного или нескольких опытных википедийцев из большого раздела было бы полезным.

Информирование о разделе

Хотя энтузиасты, иницирующие создание разделов на малых языках, стараются рассказывать о Википедии в тематических форумах, периодических изданиях и другими способами, дополнительная реклама малым разделам всегда нужна. Подготовка и распространение рекламных текстов (на русском языке, если иное невозможно или нецелесообразно) — дело, доступное любому представителю вики-сообщества, желающему помочь малому разделу.

Заключение

Несомненно, главным подспорьем для развития малых разделов может быть активное внимание к ним. Недостаточно где-то иметь хороший справочный текст, важно вовремя подсказать ссылку на него. Недостаточно быть готовым помочь, нужно вовремя предложить эту помощь. Наши новые коллеги попросту могут ещё не знать о существующих возможностях.

Следить за развитием малых разделов и участвовать в помощи им можно непосредственно или благодаря проекту «Малые разделы Википедии».

Проблемы развития чувашского раздела

Николай Плотников

Продвижение Википедии среди населения: пока активное участие в наполнении статей принимают не те, кому прочитывается лекция о полезности Википедии, а совершенно случайные люди, которые *сами* натываются на Википедию. Попытка привлечения студентов результатов не дала.

Количество статей. Плюс — раздел с большим количеством статей заметнее среди других разделов. Минусы — эти заготовки редко редактируются, при просмотре случайной страницы часто высказывают болванки статей. Сравнение в этом плане чувашского и татарского разделов.

Проблема поиска источников — слава Богу, в Чувашии полно источников информации. Газеты и журналы печатают не просто желтую прессу, есть в них и полезные материалы. По выпуску энциклопедий Чувашия находится на первых местах (выпускается очень много книг с биографиями, каждый район республики в перспективе должен иметь свою энциклопедию). Для доступа в библиотеки с фотоаппаратом ограничений нет. Проблема в малом количестве изданий на чувашском языке.

Отсутствие информации в сети. Данная проблема на данный момент хоть и решается (имеется несколько сайтов, где выкладываются газетные статьи и т.д.), но все равно это только первые шаги.

Проблема используемых терминов. Отсутствие, создание новых.

Достоверность информации в Википедии. Авторитетность источников информации и проект «Выверка статей»

Вячеслав Крахотин, Владимир Медейко

Одним из основных принципов Википедии является *проверяемость*.¹ Любая информация, включаемая в статью, должна быть подтверждена общедоступными авторитетными источниками. В то же время конечным пользователям зачастую важна не проверяемость, а *достоверность* информации.

Вопрос достоверности является важным, так как Википедия часто воспринимается как обычная энциклопедия — и поэтому размещённые в ней сведения часто принимаются на веру без должного анализа, а с другой стороны обнаружение недостоверных сведений приводит к сильному падению авторитета Википедии.

Наличие в статье Википедии недостоверной информации зачастую сопровождается нарушением проверяемости, но далеко не всегда. Наиболее характерным случаем является наличие устаревшей информации (например, многие статьи скопированы в Википедию из энциклопедического словаря Брокгауза и Ефрона без необходимой коррекции). Также недостоверность может возникать в случае, когда информация статьи основана на источниках, не обладающих должной авторитетностью. Для улучшения ситуации с достоверностью информации в статьях Википедии был создан проект выверки статей.

Выверка статей — это система проверки статей доверенными участниками, способными проконтролировать отсутствие вопиюще недостоверных, заведомо ложных или нарушающих законодательство фраз в статьях, а также отсутствие вопиющих проблем в их внешнем виде. Чтобы эта система не вредила развитию проекта, используется подход, широко распространённый в сфере разработки ПО с открытым кодом: существование текущей (рабочей) версии и стабильных (выверенной и досмотренных).

Первый опыт выверки статей позволил оценить трудозатраты, необходимые для полноценного функционирования этой системы, а также возникающие при этом проблемы. Стала очевидной нехватка в Википедии целого ряда проектов, которые должны работать параллельно с системой выверки статей. Такими проектами являются построение перечня наиболее авторитетных источников в различных областях знания, создание тематических библиографий, а также обеспечение доступа авторов Википедии к источникам.

На основе опыта выверки одним из авторов была создана классификация источников для статей научной тематики, впервые представленная в данной работе.

¹ Принцип проверяемости декларируется в «Пяти столпах» — списке основных принципов «Википедии». Хотя отдельное правило, подробно регламентирующее проверяемость, в русском разделе ещё не принято и является только проектом, в английском разделе такое правило существует.

Оценка результатов поиска семантически близких слов в Википедии

А. А. Крижановский

Под семантически близкими словами (СБС) подразумеваются слова близкие по значению, встречающиеся в одном контексте. Это могут быть синонимы (*чертог, дворец*), антонимы (*запутать, распутать*) и др.

Во многих задачах умение составить список СБС, либо сравнить слова и вычислить — какие слова ближе по значению, оказывается востребованными.

Во-первых, это так называемый «поиск по смыслу», при котором пользователь вводит слово *мобильник*, но видит страницы, содержащие другие слова, например, *мобильный телефон, сотовый* и др. Поисковая система расширила или переформулировала запрос с помощью СБС.

Во-вторых, запросно-ответные системы на этапе обработки вопроса пытаются вычислить, к какой области относится вопрос пользователя, пытаются найти похожие вопросы в базе данных. Поиск вопросов основан, в том числе, и на использовании списков СБС.

В-третьих, для выбора одного из значений многозначного слова, (например, слово *граф* может обозначать либо титул, либо математический объект) используют СБС.

В-четвёртых, есть интерес к автоматическому созданию специальных словарей — тезаурусов на основе СБС. Прелесть таких тезаурусов в том, что они строятся по тексту и могут наглядно, в виде картинки, предъявить ключевые понятия, найденные в тексте, и то, как они связаны.

В-пятых, трудоёмкая задача составления словарей синонимов (и не только синонимов) требует кропотливой работы лексикографов. Своевременную помощь оказывают поисковые алгоритмы, предлагающие списки близких по значению слов для последующего вдумчивого разбора лингвистом.

Количество научных работ, посвящённых Википедии, стремительно растёт. Осветим одну из граней этого феномена, а именно: корпус текстов Википедии обладает особой привлекательностью для поисковых алгоритмов. Вики занимает нишу между, с одной стороны, размеченными корпусами текстов, а с другой — интернет-страницами (где нет никаких надёжных подсказок для алгоритмов, кроме гиперссылок и частоты слов).

Капица П.Л. писал: «...теория — это хорошая вещь, но правильный эксперимент остаётся навсегда». Однако, чтобы провести эксперимент и оценить результаты поиска близких по значению слов нужен тестовый набор, который создан людьми вручную, а не автоматически, и поэтому считается точным.

Для английского языка такой набор есть — это 353 пары слов, в оценке которых участвовало два десятка людей. Уже более 11 метрик и алгоритмов были оценены и их можно сравнить с помощью этих данных, но только для данных на английском. Именно этот набор использовался и для оценки работы программы Synarcher в английской Википедии.

Было бы интересно оценить работу алгоритмов в русской Википедии. Предлагаю проставить оценки в тестовом наборе из русских слов и приглашаю на страницу проекта: http://ru.wikipedia.org/wiki/Участник:АКА_МВГ/Wordsim

Связность статей русской Википедии

Mashiah Davidson

Википедия является необычной энциклопедией. В отличие от бумажных изданий Википедия использует гипертекст для представления информации. Под гипертекстом в первую очередь понимается возможность быстрого перехода со статьи на статью при помощи гиперссылок. Использование гипертекста не только делает энциклопедию более цельной и удобной для восприятия, но и повышает верифицируемость информации. Переходя по ссылкам со статьи на статью, намного проще заметить различия в изложении фактов.

Важно, чтобы все статьи Википедии были связаны между собой, одно из основных правил Википедии гласит, что любая статья должна быть доступна при переходе по ссылкам с главной страницы.

Анализ связности позволяет также выявлять множество недостатков существующих статей, таких как банальное отсутствие значимости, ошибки в именовании статей, отсутствие необходимых перенаправлений, неточные ссылки, ведущие на страницы разрешения неоднозначностей. Многие изолированные статьи на деле подлежат удалению или серьёзной модификации по тем или иным не связанным со связностью причинам. При анализе связности также порой обнаруживаются странности в поведении движка MediaWiki, такие как существование статей, к которым невозможно получить доступ через Веб-интерфейс, или неполнота списка двойных перенаправлений, подлежащих спрямлению.

Два конца одной палки

При анализе связности исследуется два класса проблемных статей:

- 1) статьи, не содержащие ссылок на другие статьи, так называемые тупиковые статьи;
- 2) статьи, на которые нельзя попасть из подавляющего большинства других статей Википедии, или Категория:Википедия:Изолированные статьи, среди которых отдельно выделяются страницы-сироты, то есть, страницы, на которые вообще нет ссылок из других статей Википедии.

Предмет анализа — статьи

Анализ связности напрямую связан с вопросом о том, что считать статьями. Прежде всего, стоит заметить, что речь идёт об основном пространстве имён. Но даже в основном пространстве имеются страницы, находящиеся на особом положении, среди них

- 1) страницы перенаправления;
- 2) страницы разрешения неоднозначностей;
- 3) хронологические статьи;
- 4) всевозможные служебные списки для координации работ.

Что считать статьями, ссылки с каких страниц принимать во внимание — эти вопросы первичны при постановке вопроса о связности.

Несовершенство служебных списков

Встроенные в движок MediaWiki механизмы на поверку оказываются не вполне надёжными. Они не различают ссылки между различными пространствами имён и ссылки между статьями, не принимают в расчёт страницы разрешения

неоднозначностей, и уж, тем более, не отличают ссылки из хронологических статей и служебных списков от нормальных тематических ссылок.

Текущее состояние раздела

При текущих способах учёта ссылок, в нашем разделе около 14,5 тысяч изолированных статей, то есть, около 7% от общего числа статей. Значительная часть из них является страницами-сиротами (около 12 тысяч, или 6%). Тупиковых статей всего около тысячи (0.5% от общего числа статей). Около половины тупиковых статей одновременно являются изолированными.

Больше всего изолированных статей в категориях Персоналии по алфавиту (6.5 тыс.) и Материалы ЭСБЕ (3.7 тыс.).

Сценарии работы с изолированными статьями

1. Шаблон {{сирота}}, как средство быстрого поиска.
2. Создание перенаправлений и спрямление ссылок, идущих через страницы разрешения неоднозначностей.
3. Разматывание цепочки ссылок между изолированными статьями.
4. Поиск изолированных статей в отдельных категориях.
5. Установка ссылок из тупиковых статей и статей к викификации.

Вопросы для обсуждения на круглом столе

Как это работает?

1. Использование тулсервера и AWB.
2. Языки: SQL и unix shell.
3. Планы по улучшению.

Открытые и часто задаваемые вопросы

1. Должен ли шаблон {{сирота}} быть видимым?
2. Где следует устанавливать шаблоны, в начале статьи, или в конце?
3. Почему не используется шаблон {{rq}} с параметром linkless.
4. Почему для изолированных статей используется невидимый шаблон.
5. Как сделать названия категорий для изолированных статей более удачными.
6. Как долго бот должен выжидать, прежде чем устанавливать шаблон на новую статью.
7. Почему автоматический учёт изолированных статей не внедрён в больших языковых разделах.
8. Как улучшить существующий алгоритм поиска изолированных кластеров.
9. Какой отчёт удачнее всего использовать в качестве инструмента для ВП:П:С.
10. Следует ли учитывать ссылки со служебных списков, и нужны ли они в основном пространстве статей.
11. Как сократить число правок, производимых ботом.
12. Кто умеет рисовать?

Проблемы хронологического упорядочения статей и данных

Сватковский А.В.

При работе над Википедией требования научной точности и популярной ясности зачастую вступают в противоречие. В докладе рассматривается ряд противоречий, связанных с категорией времени.

При естественном стремлении науки к унификации форм восприятия (одна из которых - время) зачастую слабо осознается историческая изменчивость таких форм и отсутствие элементарного алгоритма их трансформации. По типу формализации категории времени можно выделить следующие эпохи:

1. Отсутствие письменных источников и наличие археологической датировки.
2. Качественный счёт (по поколениям мифических героев).
3. По годам правления царей.
4. Линейные хронологии (от основания Рима, по хиджре, китайская циклическая и другие).
5. Убыстрение темпа жизни и необходимость создания статей по дням (примерно с 1600 г.).

Предлагаемые правила как ориентиры работы:

1. Датировки заведомо недостоверные в общих статьях не указывать, но создать для взаимосвязанных рядов датировок отдельные статьи.
2. Датировки приблизительные повышать в ранге, т.е. указывать не в статье о годе, а в статьях о десятилетиях, и т.д.
3. Использовать альтернативные хронологии там, где они использовались в исторических источниках. Это позволит уточнить алгоритм пересчёта.
4. В начале текста по каждому году ввести шаблон с указанием всех эр, реально использовавшихся в тот год.

Существуют следующие основные способы упорядочения материала: в пространстве, времени, тематически. Таким образом, есть 6 возможностей иерархизации этих 3 способов:

1. тема – время – пространство. Используется чаще всего для явлений, сходных в разных странах.
2. тема – пространство – время. Используется при значительном различии по странам.
3. время – тема – пространство. Используется для хронологии при больших статьях (современность).
4. время – пространство – тема. Используется для хронологии малых статей (древность).
5. пространство – время – тема. Редко используется, вроде хронологии истории России.
6. пространство – тема – время. Часто используется в статьях по странам.

Один из принципов Википедии – ее вторичность по отношению к внешним источникам. При этом не всегда учитывается, что «информация способна порождать информацию» самим фактом своей организации. С ростом Википедии неизбежно повышение внутренней связности и, что весьма важно, возможность для написания обзорных статей с использованием лишь материалов Википедии.

Социально-политические науки в Википедии

Илья Быков

С социологической точки зрения Википедия является идеальным образом научного сообщества, как его представлял себе знаменитый американский социолог, представитель структурного-функционализма Роберт Мертон. По мнению Мертона научная деятельность держится на четырех принципах: - материальная незаинтересованность научных работников; - всеобщая доступность результатов научной деятельности; - универсальность полученных знаний; - систематическая критика.

Взгляды Мертона на социологию науки, развитие науки и научного сообщества во многом сформировались под влиянием идеологического противостояния либеральной идеологии и фашизма. Достижения немецкой науки в период между двумя мировыми войнами были бесспорны: по количеству нобелевских лауреатов Германия в то время занимала первое место. Основная задача Мертона заключалась в создании альтернативной фашистской модели научной деятельности, совместимой с либеральными ценностями. Впоследствии концепция Мертона подверглась серьезной критике. Это произошло в порядке, закрепленном им в четвертом принципе. И наиболее серьезные споры вызвал первый принцип, говорящий о материальной незаинтересованности. Тем не менее, и сегодня данная концепция является идеальной моделью, к которой, на наш взгляд, следовало бы стремиться научному сообществу в любой стране мира.

Проводя параллель между Википедией и концепцией научного сообщества Роберта Мертона, можно обнаружить массу общих моментов. Более того, Википедия предстает радикальным вариантом научного сообщества. Во-первых, научные работки все-таки получают зарплату, а также признание в той или иной форме. Википедисты как минимум не получают зарплат за свою деятельность. Во-вторых, результаты работы редакторов Википедии доступны самой широкой общественности. В-четвертых, статьи Википедии должны соответствовать наиболее устоявшимся в современном обществе стандартам, то есть представлять универсальную ценность. В-пятых, систематическая критика - это основа написания качественных статей.

Вторая часть доклада сводится к анализу современного состояния социально-политических наук в Википедии и перспективам их развития.

Перекрёстная каталогизация в Википедии

Игорь Лисицин

Обзор современных систем каталогизации

Библиотечно-энциклопедическая каталогизация

Белые страницы

Желтые страницы

Ссылки

Реализация поиска/навигации на ЭВМ

Файловые системы — Windows Vista, Unix,

Нейронные сети, “Колесо обозрения”

Облака тегов

Ключевые слова

Улучшатель понимания, созданный Болотским Владимиром Евгеньевичем (набросок на сайте xsite.spb.ru, при участии В. Н. Волковой, А. Н. Фирсова, М. Л. Ханукова, А. А. Претро, В. А. Быстрякова, Я. В. Акулай)

Древовидная категоризация – плюсы и минусы

“Divide et impera” – плюсы и минусы.

Историко-философские размышления.

Несколько веков классификации и энциклопедистов.

“Объединяй, разделяя, и властвуй”.

Самопересекающееся дерево (СД).

Обнимаем необъятное с его же помощью

Подмножества СД в разных пространствах:

Геоинформационные (ГИ) системы (ГИС)

Пример реализации 2-х корневого СД на сайте 4kg.ru.

Реализация разделов Фотографии зимнего Петербурга, Немецко-Российские культурные связи

ГИ-временная система

ГИ-временная-клиентская система для бизнес-решений

Возможность и целесообразность применения СД на Википедии

Необходимость видеть как можно большее количество статей, ссылок, фотографий подчиненных статей из т.н. "родительской статьи"

Возможность и целесообразность генерации места для синтетических статей типа "Ленин в Петербурге", "Проблемы работы Дедом Морозом в Зимбабве", "Австрийская площадь в Петербурге и венская юность Адольфа Шиккльгрубера", а также автоматизация навигации к ним с других статей Википедии.

Автоматизация интеграции Википедии с социальными сетями типа vkontakte.ru, myspace.com и пр.

Подгонка wiki-engine для этих нужд.

Деление информации в Википедии по темам: способы размежевания и викисоциальные проблемы

Иннокентий Маресин

Введение

Из чего состоит контент Википедии? Ограничиваемся регулярными статьями и категориями. Что я называю «текстом»? Википедия как граф. Связность. «Престижные» места (статьи, на которые много ссылок). Изъяны метафоры графа ссылок. Википедия находится в пространстве идей.

Зачем Википедии нужны такие исследования?

Зоны безответственности. Сортировка по качеству (выверка). Проблемы разделений/объединений; недопущение дублирования. Единообразное оформление и категоризация.

Система категорий как скелет Википедии

Идея системы категорий — направленный граф. Что это означает на самом деле, или почему недопустимы направленные циклы. Нежелательность совпадающих и пустых категорий. Изъяны метафоры направленного графа категорий: подкатегории могут сильно «уходить вбок».

Как подступиться к «пространству идей»?

Проекции. Метрические пространства. «Обычная» метрика — пространство, время, физические и т.п. величины. Ультраметрика и деревья. Произведение графов. Роль интервики в практическом упорядочении статей. А что мы вообще хотим получить?

Альтернативный (индуктивный) подход

Роль шаблонов: шаблон как маркер типа статьи; навигационные шаблоны как средство перечисления. Списки, порталы — компромисс между категориями и ссылками?

Социальные проблемы, связанные с разрастанием корпуса редакторов

А почему Википедия вообще работает? Так ли страшны конфликты? Синдром «знатока всего»: откуда он растёт? Много ли у нас редакторов? Способы координации усилий. Роль тематических проектов. Какие проекты действительно являются проектами?

Специфические проблемы различных секторов Википедии

Что проще всего классифицировать? Специалисты используют разные «языки» — в точных науках. Отсутствует «нейтральная» классификация многих явлений — в гуманитарных науках. Статьи о вымыслах. Крупные массивы «однородных объектов»: населённые пункты, фильмы, и даже персоналии.

Заключение

Операционные страницы: обзор текущего состояния. Первостепенные меры, или с чего мы начнём.

Википедия и проблемы

Г. Е. Скворцов

Википедия представляет собой наибольшее из известных собрание знаний в виде статей, посвящённых отдельным темам.

Для решения проблем, особенно крупномасштабных, нужен большой массив разных знаний, ориентированных проблемным образом. Проблемы, как правило, междисциплинарны и для своего решения требуют сочетания разделов разных наук. В Википедии собраны фрагменты современных знаний. Естественно использовать многообразие знаний Википедии для решения масштабных проблем.

Возьмём, например, проблему устойчивого развития мировой системы, отдельной страны, крупных городов. Ясно, что даже первичная постановка проблемы устойчивого и успешного развития отдельного города включает в себя широкий спектр междисциплинарных вопросов. Наряду с экономикой, техникой должны быть включены проблемы жизнедеятельности, энерго- и ресурсообеспечения, транспорта и связи, культуры и массовых коммуникаций. Практически любое звено из указанной цепи при его недостаточном решении приводит к понижению эффективности городской системы согласно принципа слабого звена (устойчивость системы определяется устойчивостью его слабого звена). Приведённый пример одной из многих проблем, как совершенно ясно, требует использования значительных объёмов знаний. Для осуществления этого собрание знаний Википедии представляется вполне подходящим.

Непосредственное привлечение фрагментов знаний Википедии для проблемной их организации требует при нынешнем состоянии Википедии значительных затрат времени и средств. Это обусловлено отсутствием должной организации гигантского материала, содержащегося в объекте нашего рассмотрения. Для плодотворности Википедии несомненной первоочередной задачей представляется организация, систематизация и проблемная ориентация знаний в Википедии. Систематизация знаний в крупном масштабе может быть осуществлена; это подтверждает опыт упорядочения знаний о природе в виде построения единого естествознания. Последнее осуществлено автором доклада в двух книгах и продемонстрировано для открывательства в книге «Технология новых знаний» (Скворцов Г.Е., Кондратьев А.Н.), готовящейся к печати.

Опыт представления технологии получения новых знаний указывает на необходимость широкого исходного поля знаний, каковым может стать Википедия, как наибольшее и открытое для пользователя собрание знаний. Авторам предлагаемой книги пришлось заниматься систематизацией больших объёмов знаний, и, следовательно, имеющийся опыт может быть расширен на крупные фрагменты Википедии, а при успешном результате и на конструкцию в целом с перспективой её прогрессивного роста.

Многоязычная Википедия: языковой помощник для физика и научно-технического переводчика

Илья Марченко

Вступление

Языковая картина научно-технического мира остаётся пёстрой и многогранной. Существует значительный интерес к тому, где искать наиболее быстрой, простой и эффективной справки в затруднительной ситуации [1][2]; как помочь физика быстро и успешно освоить иностранный язык в целях профессионального общения [1][2][3]; как помочь «физику» или «лирику» добиться максимально высокого качества научно-технического перевода [4][5][6] и каковы вообще перспективы многоязычного научного общения в условиях, когда доминирование в этой области английского языка стало полным, и в обозримом будущем необратимым [7].

Пример дискуссий, возникающих среди специалистов: можно ли, и каким образом, помочь физику (с одной стороны) и филологу (с другой стороны) не переводить «собственное значение» как "own meaning", «движение N материальных точек» как "motion N of material points" или «кулоновское взаимодействие» как "coopération de Coulomb." [1][2][5]

Википедия, играющая роль словаря?

Википедия позволяет легко перемещаться между статьями, написанными на разных языках, но посвященных одному и тому же понятию, объекту или явлению. Принцип работы межязыковых interwiki-ссылок оказывается аналогичен терминологическому словарю (читатель сравнивает названия статей и узнает перевод термина на другой язык).

У interwiki есть существенные преимущества по сравнению с бумажными или электронными словарями. Например, они позволяют перемещаться между любой парой языков (а существуют ли, например, белорусско-французские словари?) То, что interwiki могут позволить узнать переводы важных неологизмов или узкоспециальных терминов, даже не обсуждается.

Однако, по мнению автора, основной, еще не до конца раскрытый, потенциал interwiki состоит в том, что привязка одной статьи к другой (корректный перевод названия) не берётся из словаря, а обдумывается и перепроверяется многими людьми в течение долгого времени.

Пример

Как перевести с русского на английский слово «скорость»? Словарь подскажет: *speed* или *velocity*. Однако какое из двух слов нужно использовать в статье по физике?

Оказывается, что *speed* и *velocity* имеют очень тонкую разницу в смыслах, которая не всегда отражается в словарях (часто даже узкоспециальных!), однако играет принципиально важную роль. *Speed* — это «модуль» или «величина» скорости, скалярная величина (см. *speed of light*, *speed of sound*). *Velocity* — скорость как векторная величина.

Эквивалентами в польском языке могут быть слова *prędkość* (скорость как вектор) и *szybkość* (скорость как скаляр). Во французском языке существуют

термины *vitesse*, *vecteur-vitesse* и *célérité*, которые также имеют существенные, но не очевидные из словаря, отличия по смыслу.

Можно привести много примеров такого рода в разных языках. Опыт показывает, что многие дву- или многоязычные словари редко позволяют осмысленно выбрать из нескольких вариантов перевода требуемый. Нужен дальнейший поиск, который требует времени и сил.

По мнению автора, эта глобальная проблема имеет шанс быть преодоленной в Википедии. Уже в 2006—2007 г. число статей по физике в крупнейших языковых разделах достигло такой величины, что автор предпочитает начинать поиск неизвестных иноязычных эквивалентов именно в Википедии, и, сохраняя известную долю скептицизма, рекомендует это своим коллегам и ученикам.

Концепция Википедии позволяет взвешенно обсуждать и исправлять даже минимально неточные interwiki-ссылки. Впрочем, у сообщества впереди еще много работы. Например, по состоянию на 12 октября 2007 г., interwiki со статьи **ru:Скорость** (где утверждалось, что скорость — вектор) необоснованно вела на статью **en:Speed**. Аналогичная путаница наблюдалась в нескольких десятках параллельных статей о скоростях в других языковых разделах.

Почему не Викисловарь?

Число записей в Викисловаре на порядок меньше числа статей в Википедии. Викисловарь (даже в идеально доработанном виде) предполагает описательное толкование определённого термина (не названия, не составного понятия, не словосочетания).

Вместе с тем, в большинстве случаев специалисту нужно не «толкование» термина, а конкретная информация о явлении или объекте (формулы, иллюстрации, внешние ссылки), которые находятся в статье в Википедии.

В Википедии читатель видит термин в контексте, что позволяет ему мгновенно уточнить значение в непростых случаях (например, понять, чем отличаются *flow*, *stream*, *flux* и *jet*; *червеобразная* и *персистентная* цепи).

Опыт использования Википедии при изучении «научно-технических языков»

Потенциал Википедии как помощника при изучении «языков науки и техники» и при научно-техническом переводе — огромен и будет расти. Википедия позволяет не только «узнать перевод», но и сразу прочесть иноязычную статью по нужной теме.

При условии знаний по специальности и минимальных знаний в иностранном языке, это позволяет мгновенно обнаружить возможный промах при переводе (например, убедиться, что «ускорение» по-польски всё-таки не *akceleracja*, а *przyspieszenie*). Заметим, что грубые ошибки в выборе терминологии при переводе часто не связаны с пониманием или непониманием предмета и носят случайный характер.

Параллельные тексты?

Несмотря на то, что большинство связанных interwiki-ссылками статей не являются параллельными текстами, польза от сравнения таких текстов (компаративистики) может быть очень велика.

В будущем, интерес может представлять и машинное статистическое сравнение. Сервис Google Translate переводит сегодня тексты, анализируя парные документы на разных языках (документы ООН, ЕС и пр.)[8] Утверждается, что на

некоторых языковых парах он добился больших успехов. Можно предположить, что многоязычная Википедия могла бы стать неплохим текстовым источником для подобной системы.

Ожидаемые проблемы для участников Википедии

Чем больше будет узкоспециальных статей в разделах на разных языках, тем больше будет вероятность добавления некорректных (но не кажущихся таковыми) interwiki. Здесь будет нужна особая аккуратность и механизмы тщательной компетентной проверки.

Не всегда существуют однозначные переводы, и, возможно, в некоторых случаях будет целесообразно привязывать interwiki не к наиболее распространённому (или единственному имеющемуся) переводу, а к странице вроде "disambiguation".

Литература

1. D. Lisachenko. *Le français par la science : une langue étrangère enseignée par un scientifique aux scientifiques*. Actes du colloque " Apprendre une langue de spécialité : enjeux culturels et linguistiques " (École Polytechnique, Paris, 14-15 Sept. 2006).
2. И.А. Марченко, Д.А. Лисаченко. *Понимание научно-технического текста на незнакомом языке: роль знаний по специальности, мотивации и сравнительного анализа*. Материалы XXXVI Межд. фил. конф., СПбГУ, Санкт-Петербург, 12-17 марта 2007 г.
3. Étiemble. *Le jargon des sciences* (Éditions Hermann, Paris, 1966).
4. P. Alberch. *Language in contemporary science: the tool and the cultural icon*. Communications du colloque "Sciences et langues en Europe" (Paris, 14-16 Nov. 1994).
5. Д.А. Лисаченко, И.А. Марченко, И.И. Кравченко, М.Е. Соркина. *Анализ ошибок при изучении иностранного языка как языка профессиональной коммуникации*. Материалы XXXV Межд. фил. конф., СПбГУ, Санкт-Петербург, 13-18 марта 2006 г.
6. J.-M. Lévy-Leblond. *La langue tire la science*. Communications du colloque "Sciences et langues en Europe" (Paris, 14-16 Nov. 1994).
7. J. Peiffer. *Plurilinguism and scientific communication //The plurilingual European tradition as a challenge to globalizaition*. <http://www2.kokugakuin.ac.jp/ijcc/wp/global/13peiffer.html>
8. A. Tanner. *Rough, but there's little lost in Google translation*. http://www.nzherald.co.nz/section/story.cfm?c_id=5&objectid=10431378

О работе арбитражного комитета русского языкового раздела Википедии

Александр Сигачев

Брокгауз и Ефрон — вики-технологии XIX века

Evgen2

Брокгауз и Ефрон — это энциклопедия счастливого мира, стоящего на пороге ужасных потрясений и великих открытий, это энциклопедия страны, которую мы потеряли в 1917 и 1991-ом. Многие статьи энциклопедии до сих пор поражают подробностью, обилием цифр, актов, ссылок. Язык "Брокгауза" — это язык толстых и обстоятельных романов, для неторопливого и размерного чтения без рекламных пауз и прочей суеты современной жизни.

Первоисточник текста Брокгауза в интернете — сайт cultinfo. Утверждается, что это полный текст Брокгауза, однако это не так. Нет 2-томов "Россия", нет ряда статей. Часть статей и иллюстраций потерялась, почти полностью отсутствует форматирование, большая часть таблиц находится в непрезентабельном виде и т.д.

В Википедии существует следующий подход к использованию материалов Брокгауза: использовать то немногое, что еще не устарело и заменять по мере возможности на оригинальный текст, написанный участниками. Кроме того, возможно использование материалов Брокгауза по лицензионным соображениям. Альтернативный подход сохранить как есть, как историко-культурное наследие, адаптировать для современного читателя — викифицировать и разбивать текст на читаемые части, добавлять иллюстрации, ссылки, в том числе на современные материалы.