РАСПРЕДЕЛЕННАЯ ОБРАБОТКА И АНАЛИЗ ТАКСОНОМИЧЕСКИХ ДАННЫХ

В.В. Андрюсенко

Информационные технологии создают совершенно новые возможности координации деятельности системы ботанических садов по формированию богатой национальной коллекции и предоставляют каждому ботаническому саду уникальную возможность сравнительного анализа собственных коллекций с целью формирования индивидуальной коллекционной политики, направленной на повышение уникальности коллекций. Тем самым решаются задачи по сохранению фиторазнообразия, т.к. чем больше уникальность каждой коллекции, тем большее число таксонов сохраняется *ex situ*. К настоящему времени создана технологическая, методологическая и организационная база формирования информационно-поисковой системы (ИПС) «Ботанические коллекции России и сопредельных государств». В ИПС включены данные о коллекциях сосудистых растений, представленных в 76 ботанических садах, дендрологических парках и опытных станций России (Прохоров А. А. Информационные технологии для ботанических садов. [Электронный ресурс]. — Электрон. текстовые, граф., дан. (10 Мб). — Петрозаводск: 2007 (CD-ROM).

Поступающие сведения о коллекциях растений содержат достаточное количество ошибок в связи с рядом номенклатурных проблем, «таксономическим свободомыслием» отдельных кураторов крупных коллекций, отсутствием доступных источников информации по современной ботанической номенклатуре и неточностью определения растений в небольших слабофинансируемых ботанических садах. Экспертиза данных показала, что суммарный вклад использования синонимов видовых названий и ошибок в написании наименований сортов на 18% увеличивает число таксонов. Особо важной проблемой является воспроизводство ранее обнаруженных и ликвидированных ошибок при обновлении данных, поступающих из ботанических садов.

В связи с вышеизложенным, в целях предоставления корректных данных о ботанических коллекциях необходима проверка корректности научных (латинских) названий таксонов растений. Обычно такая проверка подразумевает работу со специализированной ботанической поисковой системой и ручную коррекцию названий в исходной базе данных о коллекциях. Помимо этого, требуется зафиксировать дополнительную информацию о каждой проверяемой таксономической единице.

После проверки, требуется выявить информацию о качестве конечных данных о ботанических коллекциях – количество ошибок, не поддающихся исправлению, исправленных ошибок, корректных названий, синонимов. Все эти данные показывают, насколько адекватно отражается состояние процесса сохранения биологического разнообразия, и насколько правильными будут статистические данные, полученные на основе анализа данных о ботанических коллекциях.

Для наиболее полного предоставления корректных данных требуется пользоваться не одним источником, а несколькими, данные которых перекрываются, что позволяет оценить их достоверность. Ссылки на такие ресурсы размещены нами на сайте Совета ботанических садов России [http://hortulanus.narod.ru/bgr/bgr_r.htm]. С другой стороны, такой метод сильно затрудняет процесс сбора и обработки такой информации ввиду ручного подхода к работе с источниками данных.

Ввиду рутинности всех операций по сбору и обработке таксономической информации, такой процесс довольно трудоемок для специалиста, производящего таксономическую коррекцию больших объемов таксономической информации, что в этом случае повышает риск внесения дополнительных ошибок в корректируемые данные.

Для снижения риска человеческого фактора, а также облегчения работы специалистов и предоставления им новых возможностей в их работе, нами была разработана система “Обработка номенклатурных данных о коллекциях ботанических садов”, которая автоматизирует этот рутинный процесс и позволяет перейти на более качественно новый уровень работы специалистов.

Данная система автоматически собирает и анализирует требуемые данные о таксономических коллекциях, создавая среду для работы специалистов и предоставлению им всей необходимой информации в удобном для восприятия виде. В первую очередь система используется для таксономической проверки данных о ботанических коллекциях, представленных в ИПС “Информационно-поисковая система «Ботанические коллекции России и сопредельных государств»” (http://garden.karelia.ru/).

Из наиболее полных ботанических поисковых систем, предоставляющих информацию о таксономических данных мы выбрали две наиболее крупные системы, с которыми будет производиться работа :

- 1996-2008 The International Organization for Plant Information (http://www.bgbm.org/IOPI/GPC/query.asp)
- 2004-2008 The International Plant Names Index (http://www.ipni.org)

Данные из этих систем представляют общепринятые таксономические названия и отображают детальную информации о их состоянии, авторах и источниках, в которых они были описаны.

Условно сбор и анализ данных можно разделить на три этапа :

- получение списка таксонов по каждому роду, дополнение исходного списка
- получение данных по каждому таксону и оценка этих данных
- анализ всех собранных данных совокупно (итоговый отчет)

На третьем этапе производится обработка списка таксонов, заданного специалистом. Это позволяет привлекать для проверки специалистов из различных ботанических групп – голосеменных, розоцветных и так далее. Данные специалисты в большинстве не пересекаются в своей работе и могут работать независимо друг от друга, что позволяет получать требуемые результаты в более короткий срок.

Для снижения затрат времени, а также для увеличения эффективности работы, было решено сделать распределённую обработку данных на нескольких компьютерах, имеющих выход в интернет. Основной причиной для этого служили низкие скорости отдельно взятого канала связи с интернет, а также большие объемы информации, возвращаемые источниками данных. Для этой цели были сформированы требования к функционированию системы:

- сбор требуемых данных должен обеспечиваться за приемлемое время (распределенне запросов по агентам)
- данные, полученные из источника, должны преобразовываться в структурированную форму
- для сбора данных используются агенты, запущенные на компьютерах с выходом в интернет
- собранные данные должны храниться в одном месте и предоставляться по запросу от терминала пользователя
- допускается одновременная работа нескольких терминалов со своими наборами проверяемых данных (сессиями), потенциально на компьютерах где установлены агенты
- каждый терминал должен хранить данные о текущей сессии работы (заданный список и результаты проверки)

На основе этих требований была разработана клиент-серверная архитектура, которая определила компоненты системы и уровень взаимодействий между ними :

Агент (клиентская программа сбора данных) :

- получение запроса от менеджера
- получения данных из источника на основании запроса
- первичная обработка полученных данных из источника (нормализация данных)
- выдача результирующих данных менеджеру

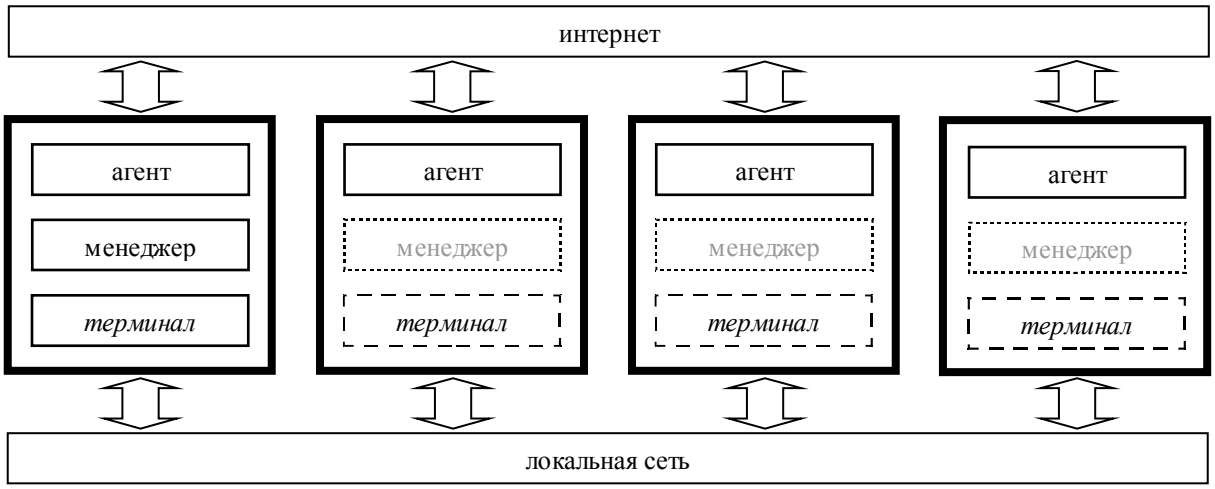
Менеджер данных (сервер) :

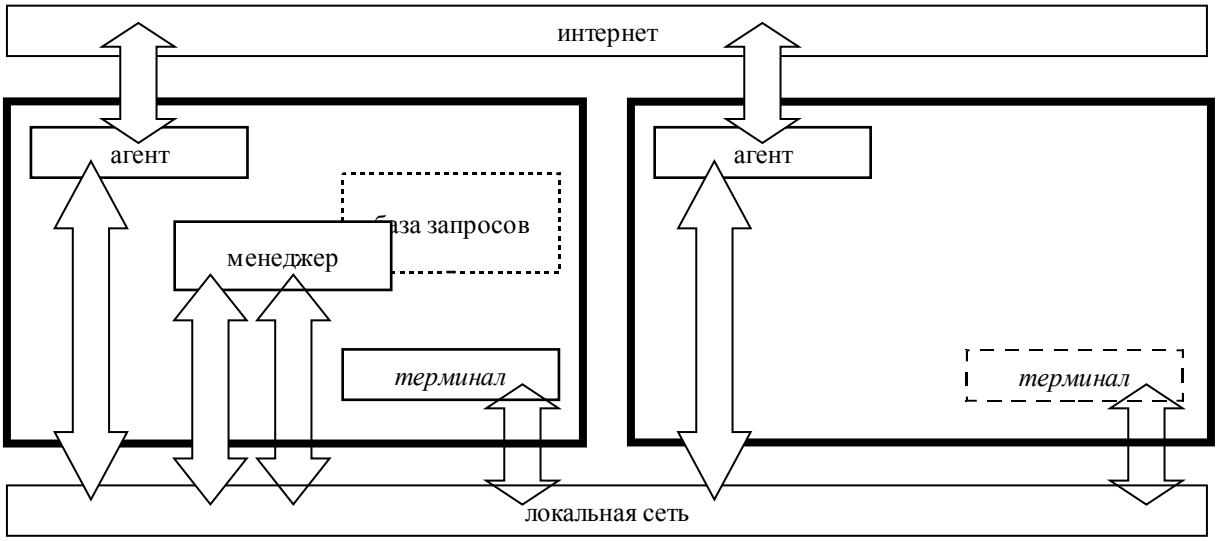
- учет активных агентов и распределение запросов между ними
- получение данных от агентов и вторичная обработка (структурирование этих данных)
- учет и хранение обработанных данных
- управление хранимыми данными
- получение запросов от терминалов и предоставление требуемых данных
- формирование отчетов по собранным данным (общая статистика)

Терминал (клиентская программа для работы специалистов)

- получение списка проверяемых таксономических данных
- формирование запросов к менеджеру на основании списка
- получение результатов запросов и регистрация статуса проверяемых таксонов
- подсчет статистических данных о состоянии проверяемых данных
- формирование отчетов на основе обработки данных (в том числе формирование чек-листов для автоматического исправления ошибок в системах хранения данных о коллекциях)

Для функционирования системы требуется группа компьютеров, объединенных локальной сетью и имеющих выход в интернет. Для установки менеджера выбирается один из компьютеров, но при необходимости менеджер может быть перенесен на другой компьютер. При необходимости, может быть установлено несколько менеджеров для разных целей (в случае работы с разными таксономическими группами или источниками данных).





После установки менеджера, на компьютеры устанавливаются агенты, которые регистрируются в менеджере, и во время ожидания периодически связываются с менеджером с целью получения запроса на сбор данных из источника. После сбора данных и выдачи результатов, агент возвращается в режим ожидания нового запроса.

Так как менеджер запоминает результаты выполненных запросов, то в случае повторения такого запроса данные выдаются немедленно, без обращения к агенту. Такая возможность снижает нагрузку на источник данных и позволяет обновлять данные при работе с похожими таксономическими списками (например, при обработке данных о коллекции конкретного ботанического сада). Для того, что бы данные отражали актуальную номенклатурную информацию о таксономических единицах, собранные данные регистрируются в менеджере, с возможностью указания периода достоверности этих данных. В случае истечения срока достоверности таких данных они обновляются автоматически.

На любой из этих компьютеров может быть установлен терминал, который обеспечивает специалисту удобный интерфейс управления процессом сбора и обработки таксономической информации. В процессе сбора информации производится отображение статистической информации о статусе проверки - сколько проверено, сколько осталось, количество ошибок, текущее положение в списке. Сессию терминала с обработанными данными можно использовать в качестве справочной базы данных для уточнения требуемой информации.

Так как система построена на основе технологии клиент-сервер, то при необходимости все три компонента могут быть запущены на одном компьютере. В этом случае теряются преимущества распределенного сбора и обработки данных, но система сохраняется свою работоспособность и при появлении новых компьютеров её всегда можно расширить.

После завершения процесса сбора и обработки таксономической информации, специалист принимает решение об исправлении ошибок и дальнейших действиях. Некоторые из ошибок могут потребовать обращения к специализированной литературе или другим источникам (в том числе на специальных сайтах).

Так же результаты проверки можно оформить в базу, который можно использовать и распространять независимо от системы. В настоящее время предложенным методом осуществлена номенклатурная экспертиза отделов *Cycadophyta, Ginkgophyta, Gnetophyta, Pinophyta* для каталога «*Gymnospermae* в ботанических садах России».

Работа осуществлена при поддержке Рособразования (проекты РНП.2.2.1.2.2308 и РНП.2.2.3.1.2306).