



## Специальность «Консолидированная информация» на кафедре информатики и интеллектуальной собственности НТУ «ХПИ»

зав.каф., проф. М.Н.Солощук  
доц. А.С.Деревянко

### 1. История и предпосылки

Кафедра информатики и интеллектуальной собственности (ИИС) была создана в Национальном техническом университете «Харьковский политехнический институт» (НТУ «ХПИ») в 1999 году на базе многолетнего опыта по переподготовки специалистов в сфере интеллектуальной собственности, накопленного в Межотраслевом институте последиplomного образования при НТУ «ХПИ». Кафедра ИИС является выпускающей кафедрой университета, обеспечивая подготовку бакалавров по направлению 6.050101 «Компьютерные науки» и магистров (специалистов) по специальности 8(7).000002 «Интеллектуальная собственность».

К настоящему времени (июль 2009 г.) кафедра уже пять раз выпускала специалистов и три раза – магистров квалификации «инженер - системный аналитик, специалист по интеллектуальной собственности» и «инженер - системный аналитик, исследователь по интеллектуальной собственности» соответственно.

Еще при создании кафедры возникал вопрос о введении для магистров специальности «Консолидированная информация», но в то время инициативная группа, создававшая кафедру, не сочла возможным принять это предложение, во-первых, из-за нехватки ресурсов, во-вторых, из-за того, что в то время мы были просто не готовы к освоению этой специальности.

Специфической областью интересов кафедры являлось в то время (как, впрочем, и сейчас) корпоративное программное обеспечение и интеграция информации. Развивая это направление в рамках подготовки бакалавров компьютерных наук, мы уделили много внимания постановке и развитию таких основополагающих направлений информационных технологий (ИТ), как базы данных, технологии Java (в том числе, и для распределенных приложений) и XML, методологии и средства проектирования и разработки программного обеспечения (IDEF, UML) и т.д. При этом в средствах ИТ основной упор делался на многоплатформенное и переносимое программное обеспечение, основанное на открытых стандартах. Наше глубокое убеждение состоит в том, что только опираясь на открытые стандарты, возможно строить информационные системы любой сложности, наилучшим образом удовлетворяющие потребности пользователей и объединяющие все лучшее, что предлагается различными производителями.

Со временем, однако, вопрос о специальности «Консолидированная информация» вновь встал, но уже не как вариант, а как необходимость. На то было несколько причин:

1. Достигнув определенной «зрелости» в преподавании упомянутых выше направлений ИТ, мы ощутили потребность их консолидировать, показать «как это все работает вместе».

2. Развивая нашу магистерскую специальность – интеллектуальную собственность, мы пришли к выводу, что проведение серьезного патентного поиска и патентно-конъюнктурных исследований требует обработки больших объемов распределенной, разнородной и разноформатной информации, то есть, ее консолидации.

3. В последние годы отрасль ИТ столкнулась с проблемой низкой окупаемости вложений, и основной путь выхода из этого кризиса состоит, как считают специалисты, в том, чтобы «ИТ развернулись лицом к конечному пользователю». Мы видим такой «поворот» в изучении методов и создании средств, позволяющих извлечь из информации знания, имеющих реальную ценность для бизнеса.

По определению ЮНЕСКО, «консолидированная информация - это открытое знание, специальным образом отобранное, проанализированное, оцененное и, возможно, ре-структурированное и переформатированное для обслуживания насущных решений, проблем и информационных нужд определенной клиентуры или социальной группы, которые иначе не в состоянии эффективно и рационально обращаться к этому знанию, потому что оно труднодоступно в его исходной форме и распределено по многим документам. Критерии отбора, оценки, реструктуризации и переупаковки этого знания определяются потенциальной клиентурой.» [<http://unesdoc.unesco.org/images/0006/000698/069802eo.pdf>]

Одно из направлений деятельности специалиста специальности «Консолидированная информация» состоит в проектировании информационных потоков, систем организации взаимодействия персонала, стратегических планов развития организации. В части технических и программных средств инструментами для выполнения задач этого направления являются персональные компьютеры (возможно, как средство доступа к сервер-центрическим вычислениям) и аналитические пакеты. Это направление деятельности *аналитика* консолидированной информации, анализирующего и интерпретирующего ее. Однако, возможно и другое направление деятельности специалиста – решение вопросов о том, как искать информацию, которой предстоит стать консолидированной, как собирать ее, хранить, управлять ею. Это направление деятельности *интегратора* консолидированной информации. Решение вопросов интеграции информации является необходимой предпосылкой для работы аналитика. Поскольку важность этих задач является несомненной, логично предположить, что существует значительный класс программных средств, обеспечивающих решение этих задач. И это действительно так. Созданием и развитием программного обеспечения, помогающего решать эти вопросы, занимаются фирмы-лидеры в сфере информационных технологий. Более того, можно без преувеличения сказать, что *программное обеспечение* для управления консолидированной информацией является передним краем развития информационных технологий и той областью, в которой сейчас разгорается наиболее острая конкурентная борьба. Поэтому мы считаем совершенно необходимым наряду с подготовкой специалистов, *использующих* консолидированную информацию, подготовку специалистов, *создающих* консолидированную информацию и *управляющих* ею с применением широкого спектра программных средств, такое создание и управление поддерживающих.

Уже в рамках специальности «Интеллектуальная собственность» мы ввели для магистров курс «Технологии и средства консолидации информации». Консолидирующая роль этого курса показана на примерной структурно-логической схеме Рис. 1.

В рамках специальности «Интеллектуальная собственность» некоторое число дипломных работ специалистов и магистров было посвящено вопросам консолидации информации и управления консолидированной информацией, например:

- Система автоматизации классификации, рубрикации и поиска информации в Internet (Евлюхин С.).
- Управление интеллектуальным капиталом в составе ERP-системы предприятия (Сапелкин П.).



Рис. 1 – Структурно-логическая схема

- Разработка программно-аналитического инструментария планирования и поддержки бизнес-процессов создания объектов интеллектуальной собственности (Канципа А.)
- Исследование и разработка моделей и средств управления политикой доступа к информационным ресурсам (Левандовская Т.)

Дальнейшее развитие этого направления привело к тому, что в 2009 г. на кафедре была лицензирована специальность «Консолидированная информация».

## 2. Классификация

Начиная заниматься вопросами консолидации информации, мы прежде всего попытались классифицировать те методы и технологии, которые составляют круг наших интересов.

*Консолидированная* - это собранная в одном месте. Мы, однако, различаем физическую консолидацию – реально собранную в одном месте информацию и логическую консолидацию – информацию, возможно, распределенную, но, с точки зрения пользователя, находящуюся в едином хранилище, имеющую общий каталог и единообразный доступ к ней. Мы трактуем название специальности именно как информацию, консолидированную логически, доступ к которой осуществляется с одного рабочего места. Таким образом, первым измерением нашей классификации является расположение информации.

Вторым измерением является структура информации. По этому измерению информацию принято подразделять на данные и контент. Исторически сложилось так, что в сфере информационных технологий под термином *данные* часто понимают структурированные данные, причем структурированные в соответствии с реляционной моделью. Неструктурированные (или имеющие структуру, отличную от реляционной) данные называют *контентом*.

Классификация и задачи, соответствующие ее разделам, показаны в левой части рис.2.

Естественно, что исторически управление информацией в информационных технологиях началось с управления консолидированными структурированными данными, и основными средствами такого управления на сегодняшний день являются СУБД. Традиционная задача СУБД - *управление транзакциями в реальном времени*

(OLTP – online transaction processing), и типовая программа курса баз данных почти исключительно ограничивается этой задачей. Однако уже давно СУБД обеспечивают решение и таких задач, как *аналитика в реальном масштабе времени, построение хранилищ данных и добыча данных (data mining)*. Индустрия СУБД имеет уже 25-летнюю историю и естественно, что именно в этих программных средствах (или в сочетании их с другими средствами) осваивается и решение новых задач, связанных с неконсолидированной и/или неструктурированной информацией.

## ИНФОРМАЦИЯ

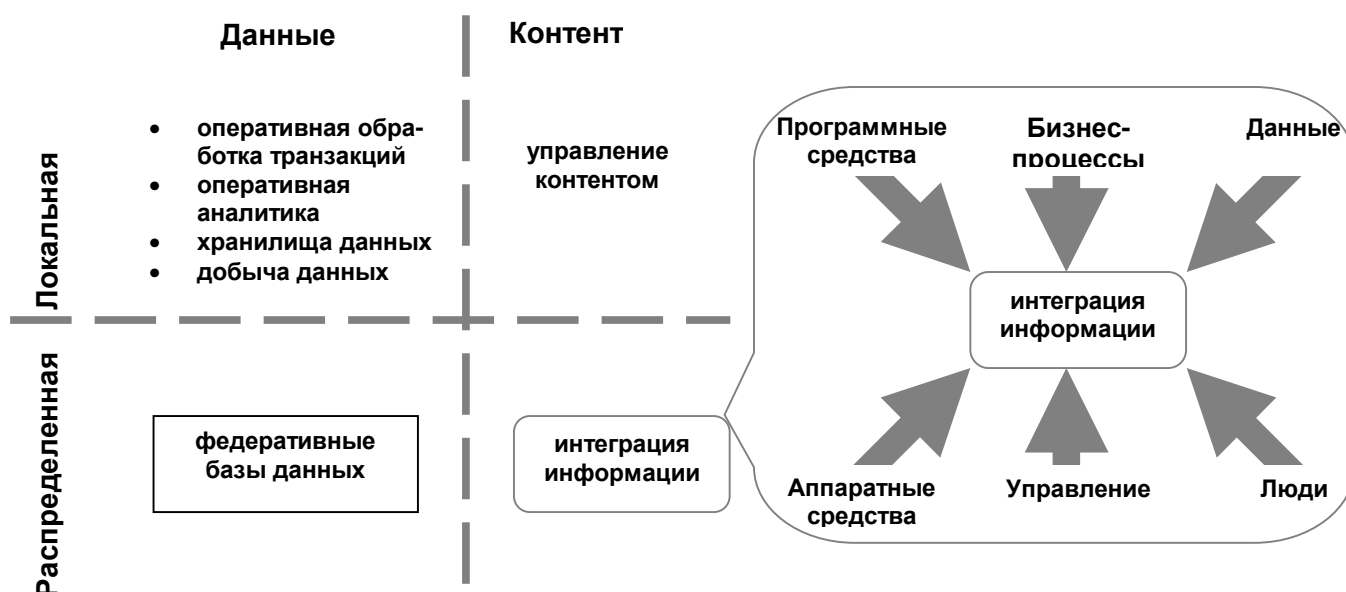


Рис. 2 – Классификация информации и задачи ее обработки

Также уже достаточно давно производители СУБД работают над решением задачи создания федеративных (распределенных) баз данных. В силу реальной распределенности предприятий, базы данных, которые обслуживают их информационные потребности, также могут быть распределены. В ряде продуктов СУБД подход, обеспечивающий такое представление распределенных структурированных данных, что для пользователя они выглядят как бы находящимися в единой локальной базе данных, называется *федеративными базами данных*.

Однако задачи, стоящие перед современными информационными системами, не позволяют им ограничиваться только данными, имеющими реляционную структуру. В современные информационные потоки включается такая информация, как аудио- и видеоданные, текстовые документы большого объема («плоские» и структурированные), документы электронной почты и т.д., и т.п. Естественно, возникает потребность обеспечения для таких документов той же богатой функциональности, которую обеспечивают СУБД для структурированной информации. Комплекс технологий, решающих эти проблемы, называют технологиями *управления контентом*, он включает в себя как технологии СУБД, так и совершенно новые подходы к построению баз данных (основанные преимущественно на технологиях XML).

Наконец, управление неструктурированной и федеративной информацией порождает задачу *интеграции информации*, которую можно считать синонимом логической консолидации. По мере развития сетевых технологий границы между локальной и распределенной информацией размываются. Поэтому нельзя говорить о технологиях, относящихся, например, только к федеративным базам данных или только

к управлению контентом. В процессе интеграции информации все технологии взаимно пересекаются и дополняют друг друга. Введенное нами разделение, возможно, в большей степени отражает потребности структурированного изложения, чем реальное разделение технологий.

Задача интеграции-консолидации не исчерпывается только вопросами управления данными/контентом. Интеграция должна быть комплексной. Интегрироваться должны бизнес-процессы, программные средства, данные, аппаратные средства, люди, работающие с информацией и, наконец, процессы управления. Составляющие комплексной интеграции информации показаны в правой части рис.2.

Интеграция *программных средств* – задача наиболее сложная, поскольку именно программные средства обеспечивают интеграцию всех остальных составляющих. Основу этой интеграции составляют серверы приложений, реализующие программные платформы, которые обеспечивают разработку и поддержку выполнения распределенных приложений. Но все большее значение приобретают средства, позволяющие объединять в одном бизнес-процессе разноплатформенные компоненты. Основу таких средств составляет сервисно-ориентированная архитектура, устанавливающая представление компонентов в виде сервисов, для потребителей которых несущественны детали аппаратной и программной платформы компонента-сервиса, языка программирования, на котором он реализован и т.д., и т.п., а важен только его интерфейс, который соответствует общепринятому открытому стандарту.

Интеграция *бизнес-процессов* состоит в приведении бизнес-процессов к компонентной структуре. То есть, бизнес-процесс должен быть представлен как управляемый поток выполнения компонентов. Каждый компонент предоставляет пользователю какую-то информацию или выполняет обработку информации. Компоненты обладают свойством повторной используемости, то есть, один и тот же компонент может использоваться в разных бизнес-процессах. В масштабах предприятия чрезвычайно важным представляется выполнение одной и той же «подзадачи» в разных бизнес-процессах одним и тем же компонентом. Это обеспечивает не только экономию средств при последующей реализации средствами информационных технологий, но и (что не менее важно) единообразие обработки в масштабе всего предприятия.

Также и интеграция данных происходит под лозунгом *«информация - это сервис»*. То есть, с точки зрения потребителя, информации не существует хранилищ информации, форматов и структур ее хранения, а существует только сервис, доставляющий ему ту информацию, которую он затребовал в том формате, который он заказал. Все детали хранения и структурирования информации скрываются под стандартной оболочкой сервиса.

Интеграция *людей* включает в себя как обеспечение доступа к интегрированной информации каждого отдельного участника бизнес-процесса, так и обеспечение совместной работы участников.

Рабочее место участника бизнес-процесса должно интегрировать доступ к необходимой для пользователя информации и средствам ее обработки, то есть, предоставлять для своего пользователя единую точку доступа к данным и функциям системы. Рабочие места такого рода строятся как порталы – мозаичные панели, содержащие в себе вложенные зоны или окна. Содержимое каждой такой зоны формируется отдельным приложением; причем обеспечивающие работу портала приложения, с одной стороны, независимы друг от друга, а с другой, могут легко обмениваться данными и синхронизировать свою работу. Наиболее развитой является технология Web-порталов, в которой панелью портала является окно Web-браузера, но интенсивно развиваются также и технологии порталов в виде «толстых» клиентов и локальных приложений.

Другой составляющей интеграции людей является обеспечение совместной работы виртуальных команд – общего документооборота, общего календаря, электронной почты, доски объявлений, виртуальных конференций и т.д., и т.п. Традиционно такие средства обеспечивались отдельными средствами, но сейчас имеется тенденция интеграции их с программным обеспечением рабочих мест и включения средств командных коммуникаций в панель портала.

Интеграция *аппаратных средств* решается в комплексе с интеграцией всех ресурсов информационной системы (аппаратуры, программного обеспечения, данных и т.д.). Основным подходом к решению этой задачи является виртуализация ресурсов, которая реализуется по двум противоположным направлениям. С одной стороны, возможно разделение ресурсов одной вычислительной системы на несколько виртуальных вычислительных систем, в каждой из которых может обеспечиваться своя операционная среда. Это концепция «консолидации серверов», позволяющая физически консолидировать на одной вычислительной системе задачи и данные, требующие разных сред. С другой стороны, возможно представление ресурсов, находящихся в разных узлах сети (возможно, глобальной сети), как единой вычислительной системы. Это концепция «вычислений по требованию», позволяющая динамически «собирать» виртуальную систему с именно такими свойствами и с такой мощностью, которые требуются для решения данной конкретной задачи.

Для предприятия с большим объемом информации и разнообразными задачами деятельности в бизнес-процессах участвует множество компонент, поддерживаемых множеством разнообразных аппаратных средств и разнообразного программного обеспечения. Поэтому чрезвычайно важной является задача единого *управления* инфраструктурой интегрированной системы. Программное обеспечение, обеспечивающее такое управление, применяет концепцию виртуализации, тесно связанную с «вычислениями по требованию», представляя в виде единой перспективы все аппаратные и программные ресурсы, участвующие в выполнении одного бизнес-процесса, независимо от того, где эти ресурсы размещены физически.

Рис.2 отражает ту классификацию, которую мы приняли, разрабатывая концепцию специальности. Эта классификация доказала свою полезность в процессе подготовки учебных планов и материалов специальности, но эта работа также показала необходимость дополнить задачу интеграции информации подзадачей *управления знаниями*. Эта подзадача включает в себя методы и средства извлечения из интегрированной информации знаний, имеющих непосредственную ценность для бизнеса и обеспечение доступности этих знаний для тех людей, которые на их основе принимают решения. Наряду с рядом технологий, рассматриваемых в других разделах классификации (OLAP, data mining) управление знаниями использует такие технологии как порталы знаний, социальные сети, синдикация новостей и, конечно, семантический Web.

### **3. Источники технологий**

По определению, консолидация требует объединения информации из разных источников, имеющей разное происхождение, разную структуру и, возможно, разное местонахождение. Поэтому ключевым моментом для консолидации является применение некоторых стандартов, если не в представлении, то в описании информации, в ее передаче, в организации ее обработки. Чтобы не ограничивать разнообразие источников информации и средств ее обработки, эти стандарты должны быть открытыми. Использование открытых стандартов дает возможность любому

производителю данных сделать поддерживаемый им источник информации доступным для извлечения из него данных любым (авторизированным) пользователем. С другой стороны, пользователь данных может обрабатывать их тем способом, который окажется наиболее приемлемым для его задачи. Открытые стандарты являются той платформой, на которой возможна честная конкуренция и взаимодействие программных продуктов от разных производителей, что опять-таки дает пользователю возможность выбирать лучшие из предложений и конфигурировать свою систему обработки информации из компонентов, наилучшим образом соответствующих его задачам.

Из открытых стандартов, которые определяют технологии, представляющие для нас наибольший интерес, следует назвать стандарты ANSI SQL и стандарты взаимодействия с базами данных ODBC и JDBC; стандарты, связанные с объектными моделями, развиваемые Object Management Group; стандарты Java Community Process; огромный комплекс стандартов World Wide Web Consortium и Organization for the Advancement of Structured Information Standards, связанных с технологиями XML.

Хотя, как мы уже отмечали, программное обеспечение консолидации информации представлено на рынке (и в спектре открытых и свободно распространяемых) программных продуктов чрезвычайно широко, мы при изложении курса ориентируемся прежде всего на концепции, технологии и продукты фирмы IBM Corp. Помимо наших личных предпочтений, это обусловлено и рядом объективных причин.

Во-первых, фирма IBM на протяжении более 10 лет является бесспорным лидером в инновациях в сфере информационных технологий. Занимая в эти годы места где-то в третьем десятке в списке крупнейших фирм мира, IBM неизменно входит в тройку лидеров по объему вложений в исследования и разработки и занимает первое место в мире по количеству получаемых патентов.

Во-вторых, интегрирующее программное обеспечение является стратегическим направлением деятельности фирмы, развиваемым ею с конца 80-х годов. Фирма IBM является на сегодня единственным производителем программного обеспечения, который представляет полный спектр продуктов по всем направлениям интеграции информации и по всем этим направлениям либо является единоличным лидером, либо входит в лидирующую группу. Ближайшими конкурентами IBM являются фирмы Oracle, Microsoft, BEA.


В-третьих, все продукты интегрирующего программного обеспечения фирмы IBM базируются на открытых стандартах. Это дает возможность надеяться на то, что, если нашему выпускнику и не доведется работать с продуктами IBM, то он будет, по крайней мере, иметь хорошую базу для понимания и освоения продуктов от других производителей.


Наконец, в четвертых, IBM представляет значительные объемы информационных и программных ресурсов для учебных заведений, чем мы и пользуемся на протяжении ряда лет. Наше сотрудничество с фирмой IBM началось еще до создания кафедры и особенно активизировалось в последние годы. Основными вехами этого сотрудничества в академической сфере явилось получение преподавателями кафедры грантов IBM Faculty Award на выполнение как учебно-методических, так и научных разработок и стажировка ряда студентов в лаборатории IBM в Цюрихе.


Почти весь спектр технологий IBM для интегрирующего программного обеспечения, укладывается в пять брендов программных продуктов.


**Information Management** Семейство продуктов *IBM управления информацией* базируется на СУБД DB2, являющейся прямой наследницей первой реляционной СУБД IBM System R. Помимо собственно СУБД, в это семейство входит целый ряд продуктов управления информацией, некоторые из которых поставляются вместе с СУБД, а некоторые – как отдельные продукты. Эти продукты решают такие задачи, как

оперативная аналитика, хранилища данных, добыча данных, управление контентом и т.д., и т.п. Следует упомянуть, что фирме IBM принадлежит также СУБД Informix, купленная несколько лет назад у одноименной фирмы. Несмотря на то, что доля этой СУБД на рынке почти в 10 раз ниже, чем доля DB2, эта СУБД имеет свои сильные стороны в некоторых технологиях. В настоящее время IBM DB2 и IBM Informix развиваются параллельно, но постоянно обмениваются технологиями, и в ближайшие несколько лет ожидается полное слияние этих продуктов.

 Семейство продуктов *IBM WebSphere* решает задачи обеспечения работы распределенных приложений и, в конечном счете, интеграции информации. Это сервер приложений, ориентированный, прежде всего, на обработку транзакций и поддерживающий как платформу J2EE, так и .NET, HTTP-сервер, сервер обмена сообщениями и др.

 Ядром семейства продуктов *IBM Lotus* являются сервер Lotus Domino и клиент Lotus Notes. Задача продуктов этого семейства – обеспечение коллективной работы. На базе Domino – Notes строятся системы документооборота, виртуальные команды, системы удаленного обеспечения, портал знаний Lotus Discovery и т.п.

 Семейство продуктов *IBM Tivoli* обеспечивает централизованное управление распределенной инфраструктурой.

 Семейство продуктов *IBM Rational* начиналось с покупки IBM фирмы Rational, разрабатывавшей средства автоматизации проектирования программного обеспечения на основе UML. В настоящее время в семейство IBM Rational входят все средства разработки IBM.

Кроме названных фирменных брендов, есть еще два или три бренда, которые принадлежат к открытым кодам, но являются стратегическими для IBM.



В широко известную операционную систему с открытым кодом Linux (<http://www.linux.org>) фирма IBM вкладывает значительные средства и передает свои технологии. Хотя продукты IBM доступны на разных платформах, именно Linux является той средой, в которой, во-первых, фирма может наиболее эффективно применять свои инновации, а во-вторых, на равных взаимодействовать с программными продуктами других производителей.



Платформа разработки приложений Eclipse (<http://www.eclipse.org>) начиналась с открытия фирмой IBM кодов своего продукта Visual Age. В настоящее время в консорциум Eclipse входят почти все «гранды» сферы информационных технологий. IBM уже не является единоличным куратором Eclipse, но продолжает оставаться одним из основных «вкладчиков» организации. Основу платформы Eclipse составляет базирующийся на Java (и, следовательно, переносимый) каркас для создания интегрированных сред разработки. Конкретные среды разработки строятся как «подключения» к ядру платформы Eclipse. Все средства разработки IBM Rational сейчас построены на платформе Eclipse. Кроме того, Eclipse предоставляет каркас для построения приложений портального типа. В 2005 году, используя этот каркас и средства семейства продуктов Lotus, IBM создала продукт IBM Workplace, представляющий собой интегрированное рабочее место, настраиваемое для любых целей.



К стратегическим направлениям следует отнести и сотрудничество IBM с организацией открытых кодов Apache Software Foundation (<http://www.apache.org>). В ходе работы над данным пособием мы неоднократно встречали примеры, с одной стороны, использования кодов Apache в продуктах фирмы IBM, а с другой стороны, – открытие фирмой кодов своих продуктов с

передачей их в Apache. Это в наибольшей степени относится к продуктам в сфере сетевых технологий, а также технологий Java и XML. Здесь также уместно упомянуть проект Apache UIMA (Unstructured Information Management Architecture), созданный на базе ядра продукта IBM Omini. Этот проект обеспечивает платформу интеграции средств поиска и добычи данных и управления знаниями на основе онтологий.

#### **4. Конкурентная разведка**

Конкурентная разведка в ее традиционном понимании, включая оперативную деятельность и специальные средства добывания информации «в стиле Джеймса Бонда», не входит непосредственно в сферу наших интересов. В большей мере нас интересует конкурентная разведка только той информации, которая может быть получена легальным путем и прежде всего из электронных источников.

С другой стороны, мы рассматриваем такие аспекты, которые связаны не только с добыванием информации, а наоборот, с представлением информации в такой форме, чтобы она была легко доступна для тех, кому она нужна. Можно сказать, что сферой наших интересов является во-первых, «конкурентная разведка в Internet», а во-вторых, «конкурентная разведка внутри предприятия/корпорации», т.е. актуализация имеющихся в Сети и на предприятии/корпорации знаний для людей, принимающих решения.

Вместе с тем, совершенно очевидно, что цели консолидации корпоративной информации и цели конкурентной разведки (как они, например, формулируются в Википедии) частично пересекаются, комплекс ИТ, участвующих в консолидации информации весьма существенно пересекается с ИТ конкурентной разведки или полностью покрывает их. Возможно, что некоторые технологии консолидации информации пока незаслуженно остаются вне поля зрения специалистов по конкурентной разведке и их освоения могут существенно обогатить последнюю.

Поэтому мы считаем, что в Обществе аналитиков и профессионалов конкурентной разведки должно быть место для специалистов разного направления: джеймсов бондов, экономистов, юристов, математиков и, конечно же, профессионалов ИТ.

Подробнее о нашем подходе к технологиям консолидации информации см. в: А.С.Деревянко, М.Н.Солощук. Технологии и средства консолидации информации. – Харьков: НТУ «ХПИ», 2008. (<http://khipi-iip.mipk.kharkiv.edu/library/extant/ii/index.html>).