InfOnCall HTML2XML

因科技术（上海）有限公司

地址：上海市淮海中路 381 号中环广场 2201-2208 室

邮编：200020

电话：021-63916988

传真：021-63915988

E-mail：contact@InfOnCall.com

www.InfOnCall.com

1. 术语

HTML2XML 模板生成器

HTML2XML 解析引擎

2. 简介

Infoncall提供了一套HTML2XML工具，用以将HTML文档自动转换为XML文档。目前主要针对以表格数据为核心（data-centric）的HTML格式文件。这是由于XML标准主要是用以精确标识所包含的数据，而有进一步应用需求的HTML文件多以含有Table的 Data-Centric文件为主。目前该工具功能主要包括：

- 提供基于XML的语言来表达如何从HTML网页获取复杂结构；
- HTML到XML声明性文档的映射，可以根据相应的解析模板自动产生XML；
- 提供可视化工具使得开发更加的迅速和便捷。

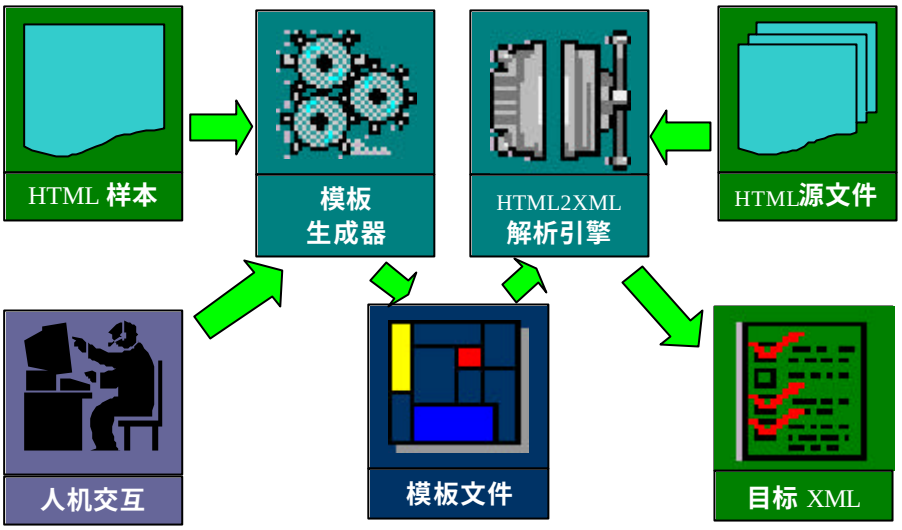
3. 背景

Internet的广泛应用和飞速发展使得以HTML表达的Web成为了信息的主要发布渠道之一。人们可以通过WWW浏览丰富的信息资源。而Web越是普及，就越迫切地要求信息内容不仅容易表现，而且能被应用方便地获取，以提供服务之间的自动化和互操作。人们要求来自Web的信息以结构化的方式来访问。W3C制定的可扩展标记语言(XML)以及其数据模型和查询语言提供了解决之道。可是如今的Web仍然是由许多杂乱的HTML网页组成，而不是组织良好的XML文档。因为需要把现有的HTML网页转换成更灵活应用和处理的XML数据。为了解决这个实际的问题，InfonCall提供了HTML2XML的开发工具，可以来将基于Web资源包装成产生所需要的XML文档。

4. 系统架构

HTML2XML1.0中包括了两个工具：HTML2XML模板生成器和HTML2XML解析引擎。通过该HTML2XML模板生成器的GUI界面和HTML2XML解析引擎，将HTML文件中的<Table>标记中的数据，根据指导

性文件，转换成XML格式数据，供其他应用程序进行进一步处理。



Infoncall的 HTML2XML 模板生成器提供方便的用户界面。HTML网页内容编辑人员，选定所需要的HTML内容后，以可视化的图形方式，用鼠标进行拖拉操作即可完成对HTML内容的获取。用户不必了解所编辑HTML文件的源代码。当保存编辑结果后，即可生成针对该类HTML文件的解析模板和DEMO解析结果。

HTML2XML解析引擎支持两种用户界面：Service和API。Service界面不需要用户有较深的编程经验；API界面为开发人员提供更灵活的编程接口。模板使用人员在开发具体应用时，通过parser 解析引擎装载不同模板，解析得到相应的结果。解析结果返回XML格式的字符串和保存为指定文件，以供进一步处理。若模板装载发生错误或开发人员未指定模板，解析引擎则按无模板的方式进行处理。此时，解析引擎解析所有Table中的数据到XML文件中。

5. 产品功能和特点

5.1 产品功能

Infoncall的HTML2XML工具，提供以下功能：

- 用户可以任意指定URL来获取Web信息；
- 目标的HTML页面可以是静态网页，也可以动态生成；

- 提供可视化的界面让用户拖拉式选择需要获取的页面元素
- 输出的方式可以是静态信息也可以是动态方式
- 可以存储、编辑和调入映射信息
- 映射规则的描述基于XML，具有扩展性

5.2 产品特点

Infoncall的HTML2XML开发工具将给您带来如下的优势：

- 有效利用现有的信息资源；
- 快速建立和商业伙伴的合作；
- 无缝升级到基于XML的网站系统
- 提供多 渠道发布的转换中间件；
- 将原有的信息的内容和表现更好的分离，有利于增加商业机会，提高企业灵活度和竞争力。

6. 应用前景

Infoncall HTML2XML工具可以应用的情景的有：

1) 网站与增值服务提供商的数据交换。

一般的情形，网站已经通过Internet发布其信息内容(比如汇率、证券信息、气象信息等)，这样的信息通常是通过其服务系统不同的格式和渠道进行发布(比如提供给WAP手机)。在进行实施过程中，要直接开放其原来的后台数据库可能对数据来源的安全性造成影响；或者有可能不同的频道信息来自不同的网站，也就可能来自不同的平台和数据库。这就需要直接针对HTML，通过调用应用服务器而不是访问后台数据库的方式来获取网页信息，并且转换成为统一的基于XML格式。XML具有独立于平台和发布渠道的特点，可以很好地用于各种不同方式的发布。

2) 网站的重新设计。

目前HTML的固有缺点已经使得原来的网站模式很难符合新的需求，特别是在商务之间相互通信的场合，XML的产生和相关技术的成熟，特别是基于XML的XHTML逐渐更新HTML，使得越来越多的网站逐渐升级到基于XML设计的网站。在这个过程中既要将新的内容以XML的方式存储和发布，同时也要考虑到兼容原来的数据。这就需要将原来的数据进行组织和转换。对

于数据库，可以通过数据库到XML的转化来实现(Infoncall也提供了通用的数据库转换到XML的工具——DB2XML)。同时许多静态的HTML网页也需要转换，其中搀杂了许多重要的信息。Infoncall HTML2XML也提供了这样机制，既可以将HTML转换成的XHTML，也可以将其转换为独立于应用的XML通用格式，然后通过XSL进行网站的发布。这将是新一代网站发展中的重要环节。