A Proposal of RSS WebCrawler Model of Product Information Service with RSS 1.0

Yusuke Nakano       Hirokazu Sagawa     Hiroyuki Tarumi
Dept. of Reliability-based Information Systems Engineering
Faculty of Engineering, Kagawa University
Kagawa, Japan
s04d455@stmail.eng.kagawa-u.ac.jp

Abstract—In this paper, we suggest a non-monopolistic distribution model of product information from publishers to subscribers by aggregators using RDF site Summary (RSS) 1.0 with extended vocabularies.

A publisher is a producer, a farmer or a food maker, etc, and submits its own product information on the web. An aggregator is a databank, a registry or a data broker, which retrieves and stores information of products. A subscriber is a seller, a distributor or a consumer, which acquires product information corresponding to the product code such as EAN/UCC Code or RFID.

The proposed model improves the quality and quantity of product information service, and also gives opportunity to enjoy new services. It gives consumers new benefits of utilizing knowledge of products, discovering unknown facts of products, etc.

Keywords- web crawler, data architecture, data mining, semantic web, database, RFID, RSS, RDF, weblog (blog), product traceability

I. INTRODUCTION

If you are going to start a food maker and sell your product to mass marketers or chain stores, they will request you to register your product information to a “databank” to synchronize their own database of EDI and POS systems. You might be also required to register data to a huge databank such as UCCnet1 (In Japan, it is JICFS/IF-DB2).

The databank distributes information from publishers like a food maker to a subscriber like a mass marketer. The more information is stored in a databank, the more efficiency a subscriber achieves in business. Therefore a subscriber chooses huge databanks like UCCnet rather than minor databanks, and a publisher registers information to huge databanks in order to make the information more reachable to subscribers. It forms a monopolistic registration model.

Because a huge databank should collect as many data as possible, data items defined in the databank cover only a comprehensive range (e.g., Brand Name, Maker Name, List Price, Size, Weight, etc.) These data items are limited to information related to sales and distribution. They are selected in order to cover most of the products. The design of database is not for consumer’s benefits. Consumers need more varieties of data items appropriate to finding various characteristics of products (e.g., Organic Foods, Additive-free Foods, Genetically Modified Organism, Allergy, Traceability, etc.)

We regard it as a monopsony of product information. Monopsony is a technical term in economics, which means a market situation in which the product or service of several sellers is sought by only one buyer. In the situation we have pointed out, the product information tends to be collected (bought) by only one huge databank.

In the situation of monopsony of product information, consumers lose chances to acquire various information they need, because collected data items are decided by only one databank rigidly.

We think the monopsony has already occurred because producers have to register data to huge databanks as we mentioned above. To avoid the monopsony and give every databank opportunities to acquire product information evenly, we propose a data distribution model in which a producer publishes its product information to the World Wide Web and a databank retrieves and stores suitable information as a contents aggregator with an RSS WebCrawler. We have adopted RSS for data description so that the aggregator (databank) can semantically analyze the collected data. In the next section, we discuss how RSS can be used for our purpose.

II. DESIGN WITH RSS

XML gives us an incentive to describe product information semantically. If a system is designed for a closed group, e.g., to implement BtoB merchandising between particular two companies, you can describe product information with XML with a specified vocabulary to be recognized semantically by a partner’s machine. However, if you open it on the web, it

---

1 http://www.uccnet.org/
cannot be recognized semantically without standardization of vocabulary with all other machines.

RDF Site Summary (RSS) 1.0 is one of standardized XML to conform to the W3C’s RDF specification. In [1], RSS is defined as follows:

*An RSS summary, at a minimum, is a document describing a "channel" consisting of URL-retrievable items. Each item consists of a title, link, and brief description. While items have traditionally been news headlines, RSS has seen much repurposing in its short existence.*

For example, when you update your personal information (e.g., announcements, ideas, journals, diaries, etc) on your Weblog site, RSS is automatically rewritten for the desktop RSS readers.

Currently Weblog tools and RSS Readers are spreading increasingly. RSS is also perhaps the most widely deployed XML application on the Web.

In this paper, we propose a publication model of product information with RSS. As a publisher called “blogger” describes personal information on Weblog sites, we expect this model based on RSS to publish product information actively.

## III. MODULARIZATION

If you need to describe something with a new vocabulary, it can be defined by extending XML-namespace and/or introducing RDF-based modularization such as Dublin Core Module.

### A. Dublin Core Module

The Dublin Core was originally developed at the March 1995 Metadata Workshop in Dublin, Ohio. The standard metadata set of Dublin Core is defined as follows [2].

<table>
<thead>
<tr>
<th>Title, Creator, Subject, Description, Publisher, Date, Language, Contributor, Source, Format, Type, Resource identifier, Relation, Coverage, Rights management</th>
</tr>
</thead>
</table>

The Dublin Core Metadata Registry (Figure 1) accumulates authoritative source of information based on above-mentioned metadata and is able to search for related vocabularies.

Similarly, we propose that new elements of the product information like the Dublin Core Module should be added to RSS and RSS descriptions based on these elements can be stored on a registry. The vocabulary of product information should be determined at the workshop like Metadata Workshop in Dublin in the future.

In the next subsection, we give a design of metadata set for product information.

### B. Three classes of product information

We have thoroughly investigated product information really used in the markets. As a result, we have found that product information can be categorized into three classes. Some examples of elements follow.

1. **Commercial Information for Dealing**

   It consists of attributes related to commercial dealings such as POS or EDI. These data items are shown below.

   | Denomination, GTIN (global Trade Item Number), GLN (Global Location Number), EAN/UCC Code, Brand Name, Maker Name, Function, Classification Category Code, List Price, Depth, Width, Height, Weight, Volume, etc. |

2. **Legally Obliged Information**

   Some data must be printed on containers or packages by law. In case of Japan, the following laws are good examples.

   - **Food Sanitation Law**
   - **The Law Concerning Standardization and Proper Labeling of Agricultural and Forestry Products**
   - **The health-promotion Law**
   - **The Law for Special Measures Concerning the Management and Relay of Information for Individual Identification of Cattle**
   - **The Law for Promotion of Sorted Collection and Recycling of Containers and Packaging**
   - **The Household Appliance Recycling Law**

---

Identify applicable sponsor/s here. (sponsors)
• Act Against Unjustifiable Premiums and Misleading Representations
• The Measurement Law
These laws impose the duty on producers or seller to show the following data items.
Food Additive, Genetically Modified Organism, Allergy, Expiration Date, Place of Origin, Calories, Sodium, Total Carbohydrates, Material, Individual Identification of Cattle, Gross Weight, Net Weight, Basic Unit, etc.

There are a lot of regulations besides the above laws. Data items that must be denoted are different in products (e.g. alcoholic beverage, tobacco, medicine, etc.) or countries.

3) Other Information
Some kind of information is provided to consumers by producers or sellers with their own intention. For example, information on traceability, food safety, farming, sales promotion, etc., is often found. Showing such information is not a duty of producers or sellers. The following data items are examples. Since there are no regulations, they can be shown in various forms: texts, images, videos, and/or sound.
Organic Farming, Grower, Agricultural Chemical, Soil and Chemical Fertilizer, Prize, etc

This class of information has many kinds of extensive items suited for the characteristic of each product.

IV. IMPLEMENTATION MODEL

To deal with the three classes of product information mentioned in section 3, we have designed a system model of product information service. The product information, which is described with RSS using vocabularies proposed in section 3, is called RSS-based Product Information in this paper.

This system model consists of three kinds of servers: RSS Registry, Ping Server, and RSS Retriever. Each server’s function is as follows (Figure 2).

RSS Registry
This is a server to which a publisher registers its own RSS-based Product Information. It sends Update Ping to the Ping Server.

Update Ping is a message that informs the change of data, defined based on “weblogUpdates.ping” opened by weblog.com (http://www.xmlrpc.com/weblogsCom). We can extend the specification if required.

Ping Server
This works to manage Update Ping received from RSS Registries, to generate “changes.xml”, and to provide it to RSS Retrievers. “changes.xml” is a description of data changes, defined based on weblog.com’s specification [3]. By using Ping Server, RSS Retrievers can aggregate RSS-based Product Information effectively.

RSS Retriever
This is an aggregation server that registers necessary tag information of RSS-based Product Information collected from distributed RSS Registry by a web crawler which synchronizes the RSS Retriever’s contents with the RSS Registry according to the received “changes.xml”.

Subscribers can access to an RSS Retriever which obtains product information. Each of RSS Retriever has its own policy or purpose. Some will deal with eggs as allergens, others will deal with "Kagawa" as place of origin. (Allergy information service will be given in the next section.)

The system works as follows:
1. A publisher registers RSS-based Product Information to a RSS Registry.
2. At the time of registration, the RSS Registry sends Update Ping to a Ping Server.
3. A Ping Server generates “changes.xml” that summarizes Update Ping received from each RSS Registry at intervals.
4. An RSS Retriever gets “changes.xml” from the Ping Server at intervals.
5. The RSS Retriever collects RSS-based Product Information that is newly registered or updated according to the received “changes.xml”.
6. The RSS Retriever takes tags that meet its own search policy. It registers these tags in the internal database from the collected RSS-based Product Information.
7. A subscriber (consumer) selects a RSS Retriever that satisfies theirs queries and gets necessary information.
V. APPLICATION EXAMPLE

As an example, we explain an application with RSS-based Product Information with a vocabulary of food allergy

A. Food Allergy

Recently, the concern about food allergy is increasing. The Japan’s Ministry of Health, Labor and Welfare announced that the specified ingredients (wheat, buckwheat, eggs, milk and peanuts) are the allergens which might cause shock and death to persons susceptible to specific allergies, and food product containing such ingredient must be indicated as containing an allergen according to the Food Sanitation Law Enforcement Regulations.

This indication corresponds to legally obliged information in section 3. In addition, the following materials as "items corresponding to specified ingredients" are not obligated but recommend indication, because they possibly cause a skin rash, a slight itching of the mouth, migraine headaches, etc.

- abalone, squid, salmon roe, shrimp/prawn, orange, crab, kiwifruit, beef, tree nuts, salmon, mackerel, soybeans, chicken, pork, Matsutake mushrooms, peaches, yams, apples, gelatine

The ministry calls on manufacturers to provide information by way of the Internet, etc. [4]

In the following subsection, we examine the description model to indicate the inclusion of the five obligated ingredients and the nineteen recommended ingredients with RSS.

B. Description Model

The number of ingredients which cause food allergies is not always single. Besides, studies of food allergies may bring new allergen to light in future. Therefore, it is desirable to describe groups of ingredients by several nodes which can be added.

W3C Working Draft provides RDF containers consisting of three types (rdf:Bag, rdf:Seq, rdf:Alt) to describe a group of resources or literals. [5]

Because the order of the ingredients is not significant, we can describe a food product containing allergens with a Bag type of RDF container.

Figure 3 is the schema which shows the product containing wheat and egg, and constructs a simple hierarchy.

In the hierarchy, a channel of product is on the top, related data of its product (title, link, description, allergy, etc) are in the second layer, and the next related data (e.g., wheat and egg) are following.

This structure can be extended with other vocabulary (Place of Origin, Calories, GMO, Agricultural Chemical, etc), and it can link to images, videos, and sound files. This model has been based on the RDF syntax.

We define the term (food:Allergy) as the root element to describe the group of allergens and the terms (wheat and egg) as common literals of the products containing these ingredients.

In this model, we propose that these elements should be described into the channel element of RSS. The model is a hierarchy of the information which puts one product on top.

C. Authentication of a Publisher

The RSS-DEV Working Group states as follows [1].

The {resource} URL of the channel element's rdf:about attribute must be unique with respect to any other rdf:about attributes in the RSS document and is a URI which identifies the channel. Most commonly, this is either the URL of the homepage being described or a URL where the RSS file can be found.
However, we propose that a product code (EAN/UCC code) should be specified on the URL of the rdf:about attribute.

Because the EAN/UCC code contains the maker code, we can identify the producer. If the address of RSS Registry Server is specified, we can also identify the publisher by the URL where the RSS file can be found. Therefore we can attest that the publisher is the real producer of the product.

D. Service for Consumers and Data-Mining

Suppose an RSS Retriever, which has a policy to aggregate information of food containing eggs with RSS-based Product Information which has been registered by food producers. If the food:Allergy element in a collected RSS data has the literal of "egg", the RSS Retriever extracts the rdf:about attribute, the title, the brief description, and food:Allergy element from the RSS data.

A consumer, as a subscriber, accesses to this RSS Retriever by a browser on a cellular phone. (s)he can perceive whether the egg is contained or not in the food and click some links to access detailed information.

When RF-tags are attached to all products in the future, you will be able to search allergens among a shopping basket collectively at a supermarket, or manage food products in the future refrigerator.

The RSS Retriever of egg allergy is supposed to be a minor databank, because most of consumers do not need it. However, for persons susceptible to specific allergies it is a much dependable databank. If the product information is supplied by only one monopolistic databank, they cannot support such 24 kinds of allergen data. Then consumers might lose a chance to acquire information like food allergy.

Furthermore, we can expect some new services using data-mining techniques. Consider a future situation where all ingredients of food are described as RSS-based Product Information and we can accumulate data of each meal at home. If one of your families becomes ill, you can access some RSS Retrievers whether some bad ingredients have been contained in past meals. Even if you cannot identify the cause of illness, your data might be dependable for other researchers. Medical researchers can collect data of meals from all patients of similar symptoms and might discover an unknown allergen with data-mining techniques.

VI. CONCLUSION

In this paper, we have proposed the RSS-based Product Information and its implementation model of information exchange. With this model, we can avoid the monopsony of product information by huge databanks. Product information would be efficiently collected by even minor databanks that has its own data aggregation policy. As such databanks can collect data from personal weblog sites, where independent writers may publish documents on their own criticisms, comments, or detailed data of products.

The RSS-based Product Information needs specific vocabularies in channel element, for various data such as food allergy. By linking channel elements, a hierarchy of product data can be constructed. In the hierarchy, data of a particular product is on the top, data of its related product is in the second layer, and the next related data follow. The RSS Retrievers and subscribers can collect data from the hierarchy as deep as possible according to their needs. This structure will enable various kinds of services to subscribers.

The RSS-based Product Information is aggregated to the RSS Retriever by RSS Web Crawlers, and synchronized to the RSS Registry by the Ping Server.

To conclude, the proposed model improves the quality and quantity of product information service, and also gives opportunity to enjoy new services. It gives consumers new benefits of utilizing knowledge of products, discovering unknown facts of products, etc.

REFERENCES