

Vaim Design Faqs

Steven Orla Kimbrough

Draft: 2006.07.09

1 What is the purpose of this document?

A vaim is a value-added information mash. The concept was introduced and essential background was given in “Vaim Faqs” (May 10, 2006), which may be found at <http://opim-sky.wharton.upenn.edu/~sok/asadai/vaim-faqs.pdf>.

In short, a vaim is a topic-oriented information system providing access to information that is amalgamated (“mashed”) from multiple sources. The original information sources are presumed to be primarily textual in nature; they are documents, not structured, numeric data.

The purpose of this document is two-fold:

1. To describe at a high level the technical design of a vaim in order concisely to communicate its capabilities to colleagues and potential users; and
2. To do so using the context of sustainability and of biofuels in particular.

2 What is the value proposition? Why should anyone want a vaim?

It is not the main purpose of this document to address this issue. Nonetheless, since it is of course the central issue with regard to the concept of a vaim, I will make some brief comments.

The “Vaim Faqs” document quotes Ellen Miller of the Sunlight Foundation (www.sunlightfoundation.com) from her blog on April 28, 2006 (<http://www.sunlightfoundation.com/node/465>). What she says is well put. She writes:

Information Mashing. Don't you just love that term? It's one of the major goals of Sunlight and while we've been working on it for the past couple of months we have a ways to go before

it happens in any substantial way. Our goal is simple: integrate in a user-friendly way individual data sets (like campaign contributions, lobbyists and government contracts) that makes the whole larger than the sum of its parts.

We'd like to create something we've dubbed an "Accountability Matrix." A website where, with one click you can look up a major donor and see not just their campaign contributions, but also their lobbying expenditures, the names of members who've flown on their private jet, the names of former congressional staffers they've hired, and so on.

In a nutshell, we want to make information more liquid and more accessible to the public.

She has politics and public policy specifically in mind. The vaim concept generalizes to any subject area of interest (e.g., sustainability) and explicitly adds the notion of adding value through leading edge software technology. The key point, however, is the generalization of what she calls the "Accountability Matrix." This is essentially a way of associating information items in a useful way. Her example is a good one. There are documents (including Federal Election Reports) about campaign contributions and there are documents (e.g., press releases, newspaper stories) about hiring of former congressional staffers. What Miller is asking for is a system that would facilitate the explicit association of (certain types of) information items.

The general principle is that fact *A and B*, that *A is associated with B*, may well be much more significant than either the fact *A* or the fact *B* in isolation. I call this the *pattern principle* or the *connecting the dots principle*. For further discussion, see our paper on pattern-oriented information retrieval [DKP00].

A vaim is a system that works to implement the pattern principle, using documents from multiple sources to make explicit associations among information items. A vaim will be useful to someone if:

1. Discovering patterns and associations implicit in the source documentation is valuable, and
2. The vaim can find, or help discover, these patterns and associations.

Regarding the second point, it must be emphasized that not all patterns and associations will be discoverable this way. Extracting information from text is difficult and fraught with noise and imprecision. In judging a vaim we shall have to examine the sorts of associations it actually can deliver. For that, details matter and it is to these details that we turn now, postponing but not dismissing the value proposition issue.

3 What are the potential information sources for a vaim?

Any obtainable collection of texts (or documents) is a potential source of information for a vaim. More specifically, and in no significant order, the following sources are realistically usable for a sustainability-biofuels vaim:

1. Patents and patent applications

US patents and registered patent applications are publicly available in text format. It is also possible to obtain and explore patents registered outside the US. However, because of the importance of the

US market the usual case is that a promising invention discovered outside the US is also patented in the US.

Patent documents contain a wealth of information about innovative products, about uses for those products, about which firms are actively developing intellectual property in a given area, and so on. In testimony to this fact, there are a number of firms specializing in extracting information from patent documents, for example ipIQ (<http://www.ipiq.com>). In no small part the value added by such firms is of an information mashing—connecting the dots—sort. For example, every patent indicates its owner, more often than not a commercial firm. It is valuable to have a database connecting firm information with patents. Companies in this space exert considerable effort in building such databases.

2. SEC filings

The US Securities and Exchange Commission, charged with overseeing the smooth functioning of securities markets, requires regular and detailed reporting from firms publicly trading on U.S. markets (<http://www.sec.gov/>). As in the case of the patent documents, these filings are publicly available in text format and are subject to examination for commercial purposes by a number of firms.

3. Company Web Sites

Nearly all firms, publicly held and traded or not, have publicly accessible Web sites, and more often than not these sites have potentially useful information. Given the URL of a company Web site it is straightforward to download all publicly available documents from the site. Typically these documents are in HTML (and so can be indexed easily) or in PDF (and so can easily be translated to text format).

Annual reports are an important type of company document. They are normally available at the company's Web site.

4. Periodicals

Any of the million or so publications with an assigned ISSN number (<http://www.issn.org/>). The most important of these will typically be collected, abstracted, and indexed by commercial firms, such as Lexis/Nexis (<http://global.lexisnexis.com/us>).

5. Government reports and archives

In the US, NTIS has a wealth of information (<http://www.ntis.gov/>):

The National Technical Information Service (NTIS) serves our nation as the largest central resource for government-funded scientific, technical, engineering, and business related information available today. Here you will find information on more than 600,000 information products covering over 350 subject areas from over 200 federal agencies.

6. Private collections and archives

The term of art for knowledge of managing private collections and archives is *records management*. The field has substantial activity and a very large annual conference (<http://www.arma.org/>).

7. General Web searches

Google and other search engine companies now offer APIs (application programming interfaces) that allow individuals to write programs that use, e.g., Google, to search the Web and retrieve documents.

4 Are there other important sources of information?

Yes, classification schemes themselves contain valuable information, and as we shall see, classification schemes combined with document collections can be effective tools for information extraction. Here is a short list of important classification schemes.

/* This list needs expanding and additional discussion for each item. */

1. USPTO: US Patent and Trade Office classification system
2. LoC: Library of Congress classification system
3. NAICS <http://www.census.gov/epcd/www/naics.html>
4. Product classification schemes
 - Manufactured and mineral products: <http://www.census.gov/prod/ec02/02numlist/02numlist.html>
 - Service products: <http://www.census.gov/eos/www/napcs/napcs.htm>UNSPC
5. A list is a simple classification scheme. Lists of companies, products, countries, cities, people, and so on will often be useful as classification systems in a vaim.

5 OK, but how does it look to the user?

There are several *access modes* by which a user can approach a vaim in order to obtain information. Briefly, they are as follows. In each case, many useful variants, features, “bells and whistles” are possible. I am leaving these out of the discussion now for the sake of focusing on the essentials.

1. Information Retrieval (IR) mode

This is the least interesting access mode, but one that readers will have encountered. The user provides the system with one or more search terms and the system returns a list of documents, ranked by putative relevance to the query. Google and other Internet search engines do this for documents on the Web.

2. Basic categorization mode

In this type of access mode the user gives the system one or more search terms *after picking a classification scheme*. The system returns lists of documents by category in the classification scheme.

For example, if the classification scheme is a list of products and the query is “heat resistant”, then the system would return lists of documents (perhaps ranked by relevance) by category: documents about heat resistance associated with shoes, documents about heat resistance associated with ships, . . . and so on for each of the categories in the classification system.

Note:

- (a) This access mode is made possible by patent pending intellectual property owned by Sizatola, LLC.
- (b) Inevitably, time will often be an important category. How query responses vary by temporal epoch will often be very informative.

3. Crosstabulation mode

This is a multi-dimensional generalization of basic categorization mode. Instead of search terms, the user specifies two (perhaps more, but that's a stretch) classification schemes and the system returns a report linking documents to the crosstab cells.

For example, if the two classification schemes were products and firms, the cells would report lists of documents about the *and* of the taxa heading the rows and columns: documents about shoes and General Motors, documents about shoes and CitiBank, documents about ships and General Motors, and so on.

Note: This access mode is made possible by patent pending intellectual property owned by Sizatola, LLC.

4. Term association mode

Term-term association frequencies have been used in a number of ways to discover information in collections of text. Most notably, Garrett Dworman in his Ph.D. thesis [Dwo99a] developed a visualization technique that affords quick and reliable discovery of information in categorized collections of documents. See also [DKP00] for a brief introduction to the method.

In term association mode, the user specifies a collection of documents, a classification scheme on which a collection has been indexed, and a search term. The system returns, by classification scheme taxa, the principal terms associated in the documents with the given search term.

Note:

- (a) Provision of this access mode may be made materially easier by, or at least combined with, patent pending intellectual property owned by Sizatola, LLC.
- (b) Term association mode assumes categorization of the underlying document collection(s). Very often, time—temporal epochs—will serve as an important categorization.

5. Information extraction mode

In this mode, the user is presented with a definite list of topics with which to query the system. These topics have been previously identified, and software and data have been developed to facilitate the queries. For example, in [CKL04] we report on an information extraction exercise in which we were able to extract uses for products described in USPTO patent documents.

Information extraction in this sense is an imperfect art, yet it may often be useful in generating usable data from collections of text.

6 How does all this work, underneath the hood?

It is not the purpose of this document to answer that question. Besides, this document is getting long enough as it is, taxing the reader's patience and concentration. Everything described in the previous section can be, and has been, done in some form or other. The important question, really a form of the value proposition question, is whether the specific things we can do in the near term will return significant value.

7 So what do you have in mind?

For the near term, features and functions that can be delivered without a huge effort. Consider a vaim focused on biofuels, motivated by sustainability considerations. There are four key architectural elements needed to support specific queries under the various access modes.

1. n-grams
2. classification schemes
3. document collections
4. mapping tables

7.1 n-grams

N-grams are words or phrases. In the near term, the vaim would be n-gram oriented in the following sense. The builders of the vaim would identify a body of n-grams and index the documents with them. As new documents appear and new n-grams are identified, indexing would be extended. Users picking search terms would (in large part) be limited to picking from the list of n-grams. The reason for this is computational tractability. Any n-gram can be used for indexing purposes, but the indexing takes time. The strategy is to identify the most important n-grams and index the collections with them. That done, queries can be completed interactively with desktop PC technology. It is not practicable at this time to index the collections fully and from that identify the n-grams.

To illustrate, Ulku Oktem has identified these “primary n-grams” in the area of sustainability:

1. Wind energy
2. Solar energy
3. Energy efficient (higher energy efficiency, energy reduction)
4. Water usage (lower water usage, water efficiency)
5. Environmentally sustainable
6. Biodegradable
7. Recyclable
8. Reduction in usage
9. Reduction in usage
10. Greenhouse gases
11. Organic certification

12. Fuel efficiency
13. Waste reduction
14. Increase productivity
15. Air pollution
16. Water pollution

This is a sample. Realistically, it should be possible to identify a few hundred n-grams of interest and use them to index the document collections. Also, the n-grams may have structural relationships among themselves, e.g., a hierarchical arrangement. Exploiting this information will be an important design goal.

Note that the n-grams serve powerfully to identify and enforce the subject focus of the vaim.

7.2 classification schemes

1. A list of firms and their URLs

May need to be assembled manually, since many of the firms of interest may be specialty firms in the biofuels space.

2. An industry classification scheme

3. One or more product classification schemes

UNSPC would serve; there are others, notably the USPTO's and the international patent classification scheme.

7.3 document collections with classifications

1. Web-posted documents associated with firms of interest

Classified by firm. Each document is downloaded from a known firm's Web site and classified as associated with that firm.

2. Available regulatory filings for the firms listed in the list of firms

SEC reports in the case of US firms. Each document is classified by its associated firm. In addition, various information, such as date, type, etc. is extracted by the documents, and placed in a relational database.

3. Patents

In the near term, only US patents. As in the case of SEC reports, each document is classified by its associated firm. In addition, various information, such as date, type, etc. is extracted by the documents, and placed in a relational database.

4. Products

This document collection would be created using the Sizatola, LLC IP. It would consist of Web documents associated with particular product categories. Of course, if another method of creating a product document base were available, it could be used too.

7.4 mapping tables

These are a series of database tables that relate data items to each. Any mashing, or amalgamation, is in large part achieved through these tables. Example:

- patents-firms, regulatory filings-firms, Web documents-firms

Most patents list a company that owns the patent. This is easily enough extracted by a program we can write. The problem is that there is no uniform or consistently used company nomenclature. So, for example, company names as they appear in SEC documents need not correspond with company names as they appear in patents. Inevitably some manual work is required. But this is necessary in order to link—mash—information from different sources. Note that using the Sizatola IP the Web document-firms table is straightforward, since we would use the list of firms to generate the searches that produce the documents.

8 Is it time to discuss the value proposition again?

Yes. Then we'll be in position to illustrate with specific queries. Key types of possible uses include:

1. Serendipities.

Surprising facts and novel associations, which are interpreted in a larger context. The blog passage by Ellen Miller, extracted above, is asking for an information system that affords exploration and serendipitous discovery. Fortune favors the prepared.

In a sense, the serendipities motivation pervades all of the potential uses of a vaim.

2. Market research.

Market research comprises a number of distinct problems and questions. Among them are the *market opportunities question*—What are the under-exploited markets for this product?—and the *new features question*—Which features, if added to this product, would greatly improve its position in various markets?

3. Product matching.

Approximately speaking, there are two kinds of products: component products and end-use products. Retailers mostly sell end-use products; industrial suppliers, component products. Component products go into the construction of end-use products. Producers buy component products in order to fabricate end-use products.

Product matching is about finding new uses for component products (this is called *product placing*) and about finding new components for an existing or envisioned end-use product (this is called *product finding*).

4. Environmental scanning.

After the French, *tour d'horizon*. This is especially apt for strategic planning purposes. The term environment scanning has become part of the received argot of business studies and, indeed, practices. The Wikipedia entry at http://en.wikipedia.org/wiki/Environmental_scanning (8 July 2006) captures much of the flavor of the concept:

For a company to gain or maintain a sustainable competitive advantage, it must be ever vigilant, watching for changes in the business environment. It must also be agile enough to alter its strategies and plans when the need arises.

Methods

There are three ways of scanning the business environment:

- Ad-hoc scanning - Short term, infrequent examinations usually initiated by a crisis
- Regular scanning - Studies done on a regular schedule (say, once a year)
- Continuous scanning - (also called continuous learning) - continuous structured data collection and processing on a broad range of environmental factors

Most commentators feel that in today's turbulent business environment the best scanning method available is continuous scanning. This allows the firm to act quickly, take advantage of opportunities before competitors do, and respond to environmental threats before significant damage is done.

There are a number of related ideas, such as SWOT (strengths, weaknesses, opportunities, and threats) analysis, many of them branded.

5. Investment analysis.

We might think of investment analysis as environmental scanning from the perspective of an investor. Environmental scanning aims to improve the position of a firm. Investment analysis—frequently undertaken at the level of an industry—aims to assess which firm(s) will be most improved, whether positively or negatively. Environmental scanning is undertaken by firm management. Investment analysis is undertaken by, or for the sake of, investors, who are presumably not committed to a particular firm.

9 OK, show me some queries you can do, and make them interesting

The emphasis here is on *some*. My aim is to provide some specific plausible ideas. A more complete roster must await a full design document.

Our framing will facilitate the presentation. We have identified five access modes (§5):

1. Information Retrieval (IR) mode
2. Basic categorization mode

3. Crosstabulation mode
4. Term association mode
5. Information extraction mode

The first, IR mode, is well-established and less interesting to us. I will make no effort to discuss it. The other modes may be matched to the five potential uses of a vaim (§8):

1. Serendipities.
2. Market research.
3. Product matching.
4. Environmental scanning.
5. Investment analysis.

It is about these potential uses that the discussion below is organized. Throughout, I shall also be concerned to indicate the mashing aspect of the examples. All of this is, insofar as possible, in the context of an envisioned vaim for sustainability and biofuels. What follows are representative examples.

9.1 Serendipities

- In *basic categorization mode* we use a *product classification scheme* indexing a collection of documents.¹ Against this categorized collection we issue queries regarding *vegetable oils* (see <http://en.wikipedia.org/wiki/Oilseed> and http://en.wikipedia.org/wiki/List_of_vegetable_oils). At the latter site we find this passage:

Major oils

Oils that account for a significant fraction of world-wide edible oil production. All are also used as fuel oils.

- Coconut oil, a cooking oil, high in saturated fats, particularly used in baking and cosmetics.
- Corn oil, one of the most common, and inexpensive cooking oils.
- Cottonseed oil, a major food oil, often used in industrial food processing.
- Canola oil/Rapeseed oil, one of the most widely used cooking oils, from a (trade-marked) cultivar of rapeseed.
- Olive oil, used in cooking, cosmetics, soaps and as a fuel for traditional oil lamps
- Palm oil, the most widely produced tropical oil. Also used to make biofuel.
- Peanut oil/Ground nut oil, mild-flavored cooking oil.
- Safflower oil, a flavorless and colorless cooking oil.
- Sesame oil, used as a cooking oil, and as a massage oil, particularly in India.

¹In doing this we employ Sizatola IP.

- Soybean oil, accounts for about half of worldwide edible oil production.
- Sunflower oil, a common cooking oil, also used to make biodiesel.

We query on each and all of these oils (coconut oil, corn oil, ...) and obtain a response measure in each case by product category. We are particularly interested product categories in which the market is large and only a small number of oils are present. If these particular oils become more expensive or even unavailable, the consequences for that product and market may be quite significant. This example is also an example relevant for each of the other four potential uses. Specifically, this could be useful for market research, for product matching, for environmental scanning, and for investment analysis.

9.2 Market research

Consider in order the two questions we identified about: market opportunities and new features. (There are of course many other questions of import in market research. These are merely examples.)

- Market opportunities.

Begin by identifying and characterizing all possibly relevant market segments, demographic groups, etc. This might be done, for example, through a subjective assessment exercise with appropriate experts, such as product managers, and consultants. Also, more than one market classification scheme could be used. For example, classifications based on age, income, and job category would presumably each be interesting. The result is a market segments classification scheme. This classification scheme may be used to create a categorized document base (using the Sizatola IP) from, for example, documents downloaded from the World Wide Web. Next a collection of n-grams would be created to cover both the principal attributes of the product to be marketed and the principal attributes that would likely characterize and distinguish the various market segments. The documents would then be indexed with the n-grams. That this point, accessing the indexed collection in *term association mode* would (likely) reveal distinguishing word pattern differences (based on the n-grams) among the various categories.

Note: This question may also be approached in the style described in §9.1.

- New features.

There are (at least) two aspects to the new features problem. First, given a new feature one can investigate who will want it and what it can be sold for. Second, given a product to which features might be added one can conceive of new features that might be assessed in the first aspect. The suggestion here is that a vaim might be useful under the second aspect. Word (n-gram) association patterns, mapped to market segments, may suggest to analysts new features that would be attractive and could be assessed under the first aspect, conventionally. Here we are in *term association mode*; this is using the vaim as an aid to creative brainstorming.

9.3 Product matching

Consider the product placing problem: Find new uses or markets for component product P .

- Product placing for product P . The salient attributes (n-grams) of P are elicited from experts and authoritative documents. These attributes are classified as pertaining to P 's properties, to uses of P , or to everything else. The attributes (n-grams) pertaining to P 's properties are used to query a document base of patents. Each 'hit', patent with matching attributes, is classified under the USPTO patent classification scheme, which is organized by product class, e.g.,

∴
 Hazardous material body cover
 Thermal body cover
 Astronaut's body cover
 Aviator's body cover
 Underwater diver's body cover
 ∴

The hits are mapped to product classes and the number of hits in each class is accumulated, producing a score for each product class, indicating its affinity with the attributes of P . In *information extraction mode* the expressed uses of the subject of each hit patent are extracted and grouped by product class. The analyst examines the identified uses (of the patented items that share many attributes with P) as potential candidates for uses of P . The original patent documents are available in each case to the analyst.

9.4 Environmental scanning

- Temporal monitoring. A list of vegetable oils is obtained (see http://en.wikipedia.org/wiki/List_of_vegetable_oils). A collection of patent documents is obtained and partitioned by year of application. For each oil in the list a report is produced showing by year of patent application: the number of patent documents in which the oil's name appears and the product taxa associated with the patent documents. Analysts examine the report in order to discern trends in uses of the oils and to suggest opportunities for uses that have not been recognized.

9.5 Investment analysis

- Crosstabulation mode: firms \times oils. Lists of firms and vegetable oils are obtained. A collection of documents is obtained (using the Sizatola II IP) in which each document is associated with at least one firm and one vegetable oil. For example, documents may be recovered from the firms' Web sites and filtered for mention of specific vegetable oils. A list of n-grams is created, indicating topics of interest. For any n-gram it is possible to provide a score of activity/relevance for the documents, by cross-category (documents linked to a particular firm and a particular oil). Analysts use these reports to help them focus on the important players in the markets for vegetable oils.
- Annotating influence diagrams. Influence diagrams are obtained from expert judgments on the key factors affecting an industry, e.g., biofuels. Among other things, these diagrams indicate directionality of influence (up or down?), and presumed wins and losses in consequence. The diagrams are driven by conditions, or events, that may or may not occur, e.g., the appearance of an economically attractive alternative market for a key input substance, which undermines the viability of products that have

relied on the substance. For example, soybeans are used to produce a number of foodstuffs, including tofu, soymilk and soy flour. Soybeans, however, may also be used to extract soybean oils, which may be used for biodiesel, a fuel that is usable by diesel engines. As the cost of petroleum-based diesel fuels rises, soybean biodiesel becomes economically attractive. Predicting the collateral effects—Who wins? Who loses?—is a prototypical investment analysis problem.

An influence diagram may be viewed for these purposes as a form of classification scheme and information may be extracted both in basic categorization mode and in crosstabulation mode. In addition term association mode may also be useful.

To take a simple example, soybeans and soybean oil may be used for a large number of purposes, including biodiesel. If the ambient market context changes greatly for any of the major uses, this will disrupt the soybean market. Further, fields that grow soybeans may instead be used to grow a number of other crops, each of which may be used for a number of purposes. If the market for any of *these* greatly changes, this could also have a strong effect on the soybean economy. We could hope to detect such changes with a vaim whose documents (and queries) were temporally classified. The patterns of query responses will change over time, signaling a potential change in a relevant market.

References

- [CKL04] Gary T. Chen, Steven Kimbrough, and Thomas Lee, *A note on automated support for product application discovery*, Proceedings of the Fourteenth Annual Workshop on Information Technologies and Systems (WITS2004) (Washington, D.C.) (Amitava Dutta and Paulo Goes, eds.), December 2004, pp. 128–133.
- [DKP00] Garrett O. Dworman, Steven O. Kimbrough, and Chuck Patch, *On pattern-directed search of archives and collections*, Journal of the American Society for Information Science **51** (2000), no. 1, 14–23.
- [Dwo99a] Garrett O. Dworman, *Pattern-oriented access to document collections*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1999, Available as a working paper, Department of Operations and Information Management, [Dwo99b].
- [Dwo99b] _____, *Pattern-oriented access to document collections*, Working paper 99-12-20, University of Pennsylvania, Department of Operations and Information Management, Philadelphia, PA, December 1999.